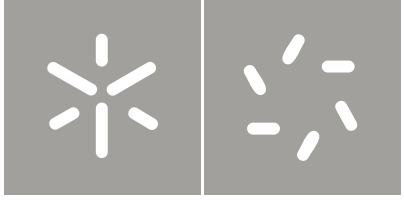
Universidade do Minho
Escola de Ciências

Alda Marisa Ribeiro Fernandes

Análise de Dados em Modelos Multiestado

Dezembro de 2010

Alda Marisa Ribeiro Fernandes Análise de Dados em Modelos Multiestado
UMinho | 2010



Universidade do Minho
Escola de Ciências

Alda Marisa Ribeiro Fernandes

Análise de Dados em Modelos Multiestado

Tese de Mestrado
2º Ciclo de Estudos de Mestrado em Estatística

Trabalho efectuado sob a orientação do
Professor Doutor Luís Filipe Meira Machado

Dezembro de 2010

Agradecimentos

É com muita satisfação que expresso aqui o mais profundo agradecimento a todos aqueles que tornaram a realização desta tese de mestrado possível.

Um especial agradecimento ao Professor Doutor Luís Meira Machado, orientador desta tese, pelo apoio, incentivo, compreensão, pelas críticas e sugestões relevantes feitas durante a orientação e pela disponibilidade demonstrada em todas as fases que conduziram à concretização deste trabalho.

Um agradecimento sincero também para os meus pais, Francisco e Conceição pelo estímulo e apoio incondicional desde a primeira hora, pela paciência e grande amizade com que sempre me ouviram, e sensatez com que sempre me ajudaram. À minha irmã Sandra que me apoiou nas mais diversas situações e por estar sempre presente. Aos meus familiares por todo apoio e carinho.

Aos amigos agradeço o companheirismo, encorajamento, confiança em mim depositada e acima de tudo a amizade.

E por fim, mas não menos importantes à Fátima pelo apoio que me deu na revisão final e ao longo deste trabalho, pela sua bondade e por tudo o que representa para mim.

Resumo

A análise de sobrevivência abrange uma classe de modelos e métodos estatísticos destinados à análise de dados de sobrevivência. Por dados de sobrevivência referimo-nos a todo o tipo de dados que resulta da observação de tempos desde um início bem especificado até à ocorrência de um evento de interesse**,** onde a 'morte' é o evento de interesse mais usual mas podendo assumir formas mais diversas como 'recaída de uma doença crónica', 'obtenção de emprego', 'execução de uma tarefa', etc.

Os modelos multiestado podem ser considerados como uma generalização da análise de sobrevivência clássica onde a 'morte' é o evento de interesse, mas onde estados intermédios são identificados. Esta modelação oferece uma ferramenta flexível para o estudo dos efeitos de variáveis preditoras (covariáveis) para as transições entre diferentes estados. Os modelos multiestado podem trazer importantes detalhes biológicos que podem ser ignorados quando se usa um modelo de sobrevivência marginal. Em contraste com os dados de sobrevivência, nestes modelos é observada uma sequência de eventos, originando mais do que uma observação por indivíduo.

Esta dissertação apresenta alguns resultados fundamentais da análise de sobrevivência. Uma revisão dos conceitos fundamentais em análise de sobrevivência, tais como, o estimador da função de sobrevivência de Kaplan-Meier, o teste de *Log-Rank* e o modelo de regressão de Cox, também conhecido por modelo de riscos proporcionais, bem como eventuais aplicações, é apresentada. A teoria básica subjacente aos modelos multiestado é também apresentada e extensivamente ilustrada. Com este trabalho pretendemos ilustrar de que modo os modelos multiestado podem ser utilizados como alternativa ao modelo de Cox.

Para demonstrar o potencial da metodologia descrita, utilizou-se a base de dados do transplante do coração de Stanford. Foram realizadas várias análises para avaliar os efeitos de várias covariáveis na sobrevivência dos pacientes, incluindo a investigação simultânea de diferentes covariáveis na sobrevivência dos pacientes a estimação do risco relativo de transplante, testes de significância e modelos de regressão. Os modelos tipo-Cox de Markov e o modelos de Markov em tempo homogéneo são apresentados, salientando-se a estimação dos efeitos de covariáveis, recorrendo ao modelo multiestado efermidade-morte. A utilização de modelos tipo-Cox revelou ser uma utilização fácil e eficiente que permite relacionar os efeitos das variáveis predictoras nas várias transições, segundo a abordagem dos modelos multiestado. Esta abordagem revelou-se uma alternativa ao modelo de regressão de Cox com covariáveis dependentes no tempo, capturando detalhes adicionais do desenvolvimento normal da doença.

Abstract

Survival analysis covers a class of models and statistical methods to analyze survival data. For survival data we refer to all types of data arising from the observation time since a well specified beginning until the occurrence of an event of interest, where 'death' is the most common event of interest, but other are also possible such as ‘relapse of a chronic illness', 'getting a job', 'execution of a task', etc.

Multi-state models can be considered as a generalization of the survival process where ‘death’ is the ultimate outcome, but where intermediate states are identified. This modeling offers a flexible tool for the study of covariate effects on the various transition rates. Multi-state models may bring out important biological insights, which may be ignored when using an ordinary survival regression model. In contrast to survival data, in these models, a sequence of events is observed, leading to more than one observation per individual.

This dissertation introduces some fundamental results in survival analysis. A review of basic concepts in survival analysis, such as the estimator of Kaplan-Meier survival function , the Log-Rank test and the Cox regression model, also known as proportional hazards model, is presented, as well as possible applications of the methods. Multi-state models framework is also presented and extensively illustrated. In this work we illustrate how multi-state models can be used as an alternative to the Cox model.

To illustrate this methodology, we have used the Stanford Heart Transplant data. A number of analyses to assess the effects of various covariates on the survival of the patients were performed, including simultaneous investigation of several covariates, the estimation of the relative risk of transplantation, significance tests and regression models. We present the Cox-type Markov model and time homogeneous Markov model focusing on the estimation of covariates effects, using the illness-death multi-state model. Cox-type regression models proved to be of easy and efficient use allowing to relate the effects of predictor variables in several transitions, according to multi-states models approach. This approach proved to be an alternative to the Cox regression model with time dependent covariates, capturing additional insights of the normal development of the disease.

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

A metodologia da Análise de Sobrevivência aplica-se a situações em que interessa medir o tempo até à ocorrência de uma determinada condição particular. Esta área estatística engloba um conjunto de métodos que se têm vindo a revelar extremamente úteis nos mais diversos campos, incluindo as ciências da saúde e da natureza, ciências sociais económicas, engenharia, entre outras. Um exemplo geralmente encontrado na literatura são os estudos clínicos, que fazem uso das técnicas de Análise de Sobrevivência para estudar o tempo de vida (ou de sobrevivência) de indivíduos e/ou avaliar o efeito de factores de risco de determinada doença.

Os modelos multiestado consistem numa generalização dos modelos de Análise de Sobrevivência clássica. Um modelo multiestado é definido a partir de um processo estocástico que permite que os indivíduos se movam entre um número finito de estados. A abordagem dos modelos multiestado permite analisar o tempo de vida de indivíduos como um processo de mudança ou transição entre estados. Deste modo, é dada especial ênfase ao progresso da doença (generalizável a outro evento) revelando os factores que afectam as diferentes transições, enquanto proporcionando simultaneamente uma visão global da doença.

Nesta dissertação apresentam-se as noções e métodos básicos da análise de sobrevivência. Introduzem-se os modelos multiestado do ponto de vista funcional, através de diagramas de forma a permitir uma melhor compreensão. Aborda-se também a questão da inferência estatística e a aplicabilidade dos modelos. É utilizada a base de dados transplante de coração de Stanford.

Este estudo está organizado da seguinte forma.

No Capítulo 2 é feita uma apresentação da terminologia própria da Análise de Sobrevivência, descrevendo-se de forma resumida os termos e conceitos básicos. Estuda-se a variável aleatória não negativa que representa o tempo de vida, ou seja, o tempo decorrido desde um instante inicial até à ocorrência de um acontecimento específico, e algumas funções muito utilizadas neste tipo de análise. Algumas distribuições paramétricas e não paramétricas para os tempos de vida são também apresentadas neste capítulo, sendo dada especial atenção ao estimador de Kaplan-Meier (1958) para estimar a curva de sobrevivência. Discute-se a comparação de distribuições de sobrevivência entre vários grupos fazendo-se uso do teste de *Log-Rank*. Neste capítulo descreve-se ainda o modelo regressão de Cox e discutem-se alguns detalhes deste modelo de regressão semi-paramétrico desenvolvido especificamente para ajustamento a dados censurados.

O Capítulo 3 apresenta os modelos multiestado. Expõe-se a metodologia destes modelos e descrevem-se os principais conceitos. Para exemplificação a abordagem é concretizada utilizando o modelo efermidade-morte (*illness-death)*. Discute-se particularmente o modelo de Markov em tempo homogéneo e o modelo de regressão tipo-Cox.

No Capítulo 4 apresenta-se um breve estudo da metodologia básica de análise de sobrevivência e dos modelos descritos. A base de dados de transplante do coração de Stanford é utilizada. Numa primeira fase efectua-se uma breve análise descritiva, calcula-se, interpreta-se e compara-se curvas de sobrevivência e realiza-se uma análise de sobrevivência simples utilizando o modelo de regressão de Cox. Numa fase posterior ajusta-se um modelo multiestado aos dados, e apresentam-se e comparam-se os resultados obtidos. Para efectuar as diversas análises utiliza-se o software R com o auxílio de várias *packages tdc.msm*, *p3state.msm* e *etm* desenvolvidas por vários pelos autores, e o software estatístico *BayesX*.

Por último, no Capítulo 5 enumeram-se as principais conclusões do estudo realizado e sugerem-se caminhos para trabalhos futuros.

Capítulo 2

Noções Básicas de Análise de Sobrevivência

2.1. Introdução

A abordagem de problemas de Análise de Sobrevivência teve início no século XVII, com o aparecimento de tabelas de mortalidade. Tendo como objectivo o estudo de dados de sobrevivência, esta técnica estatística tem vindo a tomar uma posição de destaque nas últimas décadas, quer pelo elevado número de métodos desenvolvidos, quer pelas diversas áreas onde tem aplicabilidade. São exemplos dessas áreas a Psicologia, a Demografia, a Física, a Engenharia e a Medicina.

Os dados de sobrevivência surgem quando, para um determinado grupo de indivíduos, se pretende estudar o tempo decorrido desde um instante inicial, bem definido, até à ocorrência de um acontecimento (evento) de interesse. Uma característica deste tipo de dados é que não se pode assumir a normalidade da distribuição subjacente e, por conseguinte, muitos métodos estatísticos usuais não podem ser utilizados. Adicionalmente, em muitas situações, o tempo de sobrevivência de alguns indivíduos não é conhecido com exactidão. De facto, a possibilidade de existência de dados censurados, que ocorrem quando, para alguns indivíduos, a realização do acontecimento de interesse não é observada durante o período de observação definido, distingue a Análise de Sobrevivência de outras metodologias.

A presença de censura exige, portanto, a utilização de métodos específicos, adequados a tais situações. Frequentemente, são também registadas para cada indivíduo em estudo os valores de certas variáveis, designadas por variáveis explicativas ou covariáveis, que representam factores que se supõe afectarem o tempo de sobrevivência, interpretado como variável resposta ou dependente. É de salientar que o tempo de sobrevivência de cada indivíduo tem um sentido muito vasto, pois apesar de frequentemente representar o tempo até à morte, pode também representar, por exemplo, o tempo até ao aparecimento de metástases, até à recaída ou até à cura.

A Análise de Sobrevivência tem como principais objectivos: estudar a distribuição do tempo de sobrevivência; comparar as distribuições dos tempos de sobrevivência correspondentes a dois ou mais grupos de indivíduos; modelar e inferir a relação entre o tempo de sobrevivência e variáveis explicativas ou covariáveis.

Neste capítulo introduzem-se alguns conceitos básicos relacionados com as funções mais usadas em Análise de Sobrevivência.

2.2. Censura

Em muitos estudos, nomeadamente naqueles em que existem longos períodos de *follow-up*, muitos indivíduos não atingem o tempo total de seguimento previsto. Na análise estatística clássica, dado que estes indivíduos não estiveram todo o tempo em observação, têm de ser excluídos da análise, já que se desconhece o tempo até ao evento de interesse. Estas observações podem ocorrer porque os indivíduos abandonaram o estudo, foram perdidos no seguimento ou porque o estudo chegou ao seu término sem que tivesse observado o evento de interesse. Nos estudos de análise de sobrevivência, os dados destes indivíduos são incluídos na análise, considerando-se que esses indivíduos constituem observações censuradas. Assim censura significa que para determinado indivíduo não se observou o evento de interesse, durante o período de observação do estudo (independentemente do motivo pelo qual tal aconteceu ou do que o que lhe possa ter acontecido posteriormente) (Botelho & Cruz, 2009). A grande vantagem da análise de sobrevivência é permitir incorporar a informação de todos os indivíduos na análise, através da observação do evento de interesse ou da inclusão de uma observação censurada.

Consideremos uma amostra de n indivíduos. Seja $C_i, 1 \leq i \leq n,$ o tempo potencial de observação para cada indivíduo. Seja y_i $(i = 1, \dots, n)$ a variável aleatória que representa o tempo de vida do i-ésimo indivíduo, que é totalmente observado (completo) se e só se $y_i \leq C_i$. Para cada indivíduo, as observações são da forma (T_i, δ_i), onde $T_i = \min(y_i, C_i)$ e δ_i indica a presença ou ausência de censura,

$$\delta_i = \begin{cases} 1 & se \;\; y_i \; é \; o \; tempo \; de \; vida \;\; (y_i \leq C_i) \\ 0 & se \;\; y_i \; é \; o \; tempo \; de \; censura \;\; (y_i > C_i) \end{cases}$$

Note-se que *a priori* uma observação censurada é distinta de uma omissa, já que a ordem de uma observação censurada relativamente a outra não censurada pode ser conhecida e como tal transmitir informação sobre a distribuição amostral.

Dado que existem diversos motivos para o aparecimento de dados censurados, podem ser considerados vários tipos de censura.

A censura à direita, é a mais comum em sobrevivência, ela ocorre quando, no momento em que termina o estudo, existem indivíduos para os quais não se conhece o instante exacto em que ocorreu o evento. Somente se conhece que terá sido posterior ao final do estudo. Neste tipo de censura, o valor exacto do tempo de vida será superior ao valor observado. O mesmo ocorre quando não se pode observar o momento exacto do evento de interesse devido a perda de seguimento do indivíduo em estudo. Um exemplo deste tipo de censura pode resultar de experiências clínicas que terminam num instante pré-estabelecido, portanto antes que todos os indivíduos tenham observado o evento de interesse.

Outro tipo de censura é chamado de censura à esquerda. Ocorre quando o tempo de sobrevivência é menor do que o tempo observado. Se por exemplo o acontecimento de interesse é o desenvolvimento de um tumor, os indivíduos que apresentam metástases dão origem a observações censuradas à esquerda.

A censura intervalar, ocorre quando não conhecemos o momento exacto em que ocorreu o evento de interesse, mas sabemos que este ocorreu dentro de um determinado intervalo de tempo. Este tipo de censura é frequente em situações de ensaios clínicos em que os doentes efectuam visitas regulares ao médico.

Note-se que quer a censura à direita quer a censura à esquerda são casos particulares de censura intervalar.

Como o tipo de censura mais frequentemente encontrado em Análise de Sobrevivência é a censura à direita, a partir de agora, sempre que utilizarmos o termo censura (dados censurados) estaremos a referir-nos a este tipo de censura.
Há vários mecanismos de censura (à direita), sendo os mais usuais os que se irão referenciar de seguida.

2.2.1. Censura do Tipo I

Este tipo de censura é a mais habitual em estudos médicos. É aquela em que o estudo é encerrado após um período pré-determinado de tempo, ou seja, existe um tempo t_c a partir do qual todos os dados são censurados. Se o acontecimento de interesse ocorrer antes de t_c podemos observar o verdadeiro tempo de vida do indivíduo. Caso isso não aconteça temos as chamadas observações censuradas.

2.2.2. Censura do Tipo II

É aquela em que o estudo é encerrado após ter ocorrido o evento de interesse de um número pré-determinado de n indivíduos.

2.2.3. Censura Aleatória

Quando a censura não está relacionada com o tempo de sobrevivência e ocorre de forma aleatória. O indivíduo é retirado do estudo por uma causa alheia ao próprio estudo. As razões para considerar este último tipo de censura prendem-se com a chegada ao fim do estudo ou qualquer motivo, não relacionado com o tempo de sobrevivência, que impossibilita a recolha de mais informação.

2.3. Truncatura

Uma outra característica presente em alguns estudos de análise de sobrevivência, muitas vezes confundida com a censura, é a existência de truncatura. Esta ocorre devido a um processo de selecção inerente ao planeamento do estudo. A truncatura é caracterizada por excluir os indivíduos que não são relevantes para o estudo em questão.

2.3.1. Truncatura à Direita

Este tipo de truncatura ocorre quando apenas são considerados na análise estatística os indivíduos para os quais se observou um determinado acontecimento de interesse durante o período em estudo.

Truncatura à direita, ocorre quando os indivíduos não são acompanhados a partir do tempo inicial, mas somente após experimentarem um certo evento. É particularmente importante em estudos de HIV.

2.3.2. Truncatura à Esquerda

É o tipo de truncatura mais comum e ocorre quando apenas são incluídos na amostra indivíduos que sobrevivem tempo suficiente para que ocorra um determinado acontecimento antes do evento de interesse. Um exemplo deste tipo de truncatura pode ser encontrado em estudo de sobrevivência de doentes com leucemia. Por exemplo, pacientes que foram sujeitos a transplantes de medula, e apenas os que sofrem recaída da doença são observados pelo investigador.

É de realçar que o facto de existir censura não impede a existência em simultâneo de truncatura, sendo mesmo muito usual deparar com dados que são censurados à direita e truncados à esquerda.

2.4. Distribuição dos Tempos de Sobrevivência

Considere-se T uma variável aleatória contínua não negativa que representa o tempo de vida de um indivíduo proveniente de uma população homogénea.
Suponhamos, que os indivíduos não diferem entre si relativamente a factores susceptíveis de influenciar o seu tempo de sobrevivência. Sendo assim, não iremos, por agora, introduzir quaisquer covariáveis nas definições a seguir apresentadas.
A variável T é essencialmente caracterizada por três funções, a função densidade de probabilidade, $f(t)$, a função de sobrevivência, $S(t)$ e a função *hazard*, $h(t)$ (ou função risco), ilustrando cada uma diferentes aspectos dos dados.

2.4.1. Função Distribuição

Seja $F(t)$ a função distribuição de T, definida como a probabilidade de observar o evento de interesse no intervalo $[0, t]$ e representada da seguinte forma:

$$F(t) = P(T \leq t), \qquad t \geq 0$$

$F(t)$ é uma função:

- monótona não decrescente
- contínua à direita
- $F(t) = 0$ para $t = 0$ e $F(t) = 1$ quando $t \to +\infty$.

2.4.2. Função de Sobrevivência

A função de sobrevivência representa a probabilidade de para um indivíduo não se ter observado o evento de interesse pelo menos até ao instante t.
Define-se a função de sobrevivência, $S(t)$, da variável aleatória T por

$$S(t) = P(T > t) = 1 - F(t), \quad t \geq 0$$

onde $F(t)$ é a função distribuição correspondente.

A função de sobrevivência goza das seguintes propriedades:

- monótona não crescente
- contínua à esquerda
- $S(t) = 1$ para $t = 0$ e $S(t) = 0$ para $t \to +\infty$. O gráfico de $S(t)$ é chamado curva de sobrevivência.

2.4.3. Função Densidade de Probabilidade

A função densidade de probabilidade é, por definição:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, \quad t \geq 0$$

O gráfico de $f(t)$ é designado por curva de densidade.

Uma questão frequente na Análise de Sobrevivência é, sabendo que ainda não se observou o evento de interesse até determinado instante, qual a probabilidade deste ocorrer no instante seguinte, isto é, qual a taxa instantânea de falha. A resposta é-nos dada pela função *hazard*.

2.4.4. Função *Hazard*

A função *hazard* ou função risco representa a taxa instantânea de falha no instante t, condicionada à sobrevivência até esse instante. A função *hazard* é dada por

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \quad t \geq 0.$$

A função *hazard* verifica as seguintes propriedades:

- $h(t) \geq 0$
- $\int_0^{+\infty} h(t)dt = +\infty$.

As condições anteriores são não só necessárias mas também suficientes para que qualquer função real seja função *hazard*. De facto, a definição de função hazard pode ser estendida a qualquer variável aleatória, não necessariamente positiva.
Define-se ainda a função *hazard* cumulativa ou *hazard* integrada como sendo

$$H(t) = \int_0^t h(u)du\,, \quad t \geq 0.$$

Esta função mede o risco de ocorrência do acontecimento de interesse até ao instante t, sendo uma função monótona não decrescente.

2.4.5. Relações entre as Funções

Utilizando as definições anteriores, podemos deduzir várias relações entre as funções $f(t), S(t)$, $h(t)$ e $H(t)$ em particular :

- $h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}$
- $S(t) = \exp\big(-H(t)\big)$
- $f(t) = h(t)\exp\left(-\int_0^t h(u)du\right)$
- $H(t) = -\log(S(t)$

Das relações apresentadas podemos concluir que a distribuição de T fica univocamente determinada a partir de qualquer uma das seguintes funções: função distribuição, função densidade de probabilidade, função sobrevivência, função *hazard* e função *hazard* cumulativa.

2.5. Breve Estudo de Modelos Paramétricos

Os métodos tradicionalmente utilizados na análise de sobrevivência podem-se dividir em três grupos: paramétricos, não paramétricos e semiparamétricos. Os mais usuais são os semiparamétricos e os paramétricos e as estimativas que se obtêm com estes métodos servem de apoio a posteriores analises estatísticas mais detalhadas. O processo de estimação da função de sobrevivência varia consoante o modelo (método) que se pretende ajustar aos dados.

2.5.1. Algumas Distribuições Utilizadas em Análise de Sobrevivência

Aos tempos de vida podem ser atribuídas formas específicas, isto é, podem ser modelados segundo uma distribuição conhecida. Um dos objectivos de estudos de Análise de Sobrevivência é escolher uma distribuição apropriada para se trabalhar.
Apresentam-se de seguida algumas das distribuições mais utilizadas em análise de sobrevivência.

2.5.1.1. Distribuição Exponencial

Considere-se T uma variável aleatória com distribuição exponencial de parâmetro $\lambda > 0$, com função densidade de probabilidade

$$f(t) = \lambda \exp(-\lambda t), \quad t \geq 0.$$

Então, a função de sobrevivência é dada por

$$S(t) = \exp(-\lambda t), \quad t \geq 0$$

e função *hazard*

$$h(t) = \lambda, \ \ t \geq 0.$$

Como se pode observar, a função *hazard* é constante ao longo do tempo, o que reflecte a 'ausência de memória' da distribuição. Esta importante propriedade é apenas verificada no modelo exponencial, o que faz dele um modelo bastante simples.

2.5.1.2. Distribuição Weibull

Na prática, a hipótese de função *hazard* constante é pouco realista originando assim a necessidade de outro tipo de distribuições.

Seja T uma variável aleatória com distribuição Weibull de parâmetros λ e α, isto é, $T \sim W(\lambda, \alpha)$, sendo $\lambda > 0$ o parâmetro de escala e $\alpha > 0$ o parâmetro de forma. A distribuição Weibull é definida por

$$f(t) = \lambda \alpha t^{\alpha-1} \exp(-\lambda t^{\alpha}), \ \ t \geq 0,$$

$$S(t) = \exp(-\lambda t^{\alpha}), \ \ t \geq 0,$$

$$h(t) = \lambda \alpha t^{\alpha-1}.$$

A distribuição de Weibull aparece como generalização da distribuição exponencial e tem uma aplicação muito mais vasta pois a sua função *hazard* tanto pode ser constante ($\alpha = 1$), monótona crescente ($\alpha > 1$) ou monótona decrescente ($0 < \alpha < 1$).
Esta distribuição é provavelmente a mais utilizada em Análise de Sobrevivência, nomeadamente em aplicações biomédicas e em experiencias com animais de laboratório.

2.5.1.3. Distribuição Gama

Como já se referiu anteriormente, a distribuição Weibull é uma generalização da distribuição exponencial, outra generalização da distribuição exponencial é a distribuição gama.

Seja T uma variável com distribuição Gama com parâmetros de escala e forma λ e k respectivamente. As funções densidade de probabilidade, de sobrevivência e $hazard$ são dadas por

$$f(t) = \frac{\lambda^k}{\Gamma(k)} t^{k-1} \exp(-\lambda t)$$

$$S(t) = 1 - \mathrm{I}(\lambda t, k)$$

$$h(t) = \frac{\frac{\lambda^k}{\Gamma(k)} t^{k-1} \exp(-\lambda t)}{1 - \mathrm{I}(\lambda t, k)}$$

onde $t \geq 0, \lambda > 0$ e $k > 0$. $\Gamma(k)$ é a função gama, e $\mathrm{I}(\lambda t, k) = \frac{1}{\Gamma(k)} \int_0^t u^{k-1} e^{-u} du$ é conhecida como a função gama incompleta.

A função *hazard* é constante quando $(k = 1)$, monótona crescente quando $(k > 1)$ e monótona decrescente $(0 < k < 1)$.

As distribuições Weibull e Gama têm propriedades semelhantes, no entanto esta última é menos utilizada, pelo facto de depender da função gama incompleta, tornando a especificações da função de sobrevivência e de *hazard* mais difícil.

2.5.1.4. Distribuição Log-Normal

Uma variável aleatória T tem distribuição log-normal de parâmetros μ e σ se $Y = \ln T$ tem distribuição Normal com valor médio μ e variância σ^2. Assim a função densidade de probabilidade de T é, para $t \geq 0$,

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} exp\left[-\frac{1}{2}\left(\frac{lnt - \mu}{\sigma}\right)^2\right]$$

a função de sobrevivência é dada por

$$S(t) = 1 - \Phi\left(\frac{lnt - \mu}{\sigma}\right)$$

A sua função de *hazard* é dada por

$$h(t) = \frac{\frac{1}{\sqrt{2\pi}\sigma t} exp\left[-\frac{1}{2}\left(\frac{lnt-\mu}{\sigma}\right)^2\right]}{1-\Phi\left(\frac{lnt-\mu}{\sigma}\right)}$$

onde $\mu \epsilon \mathcal{R}$, $\sigma > 0$ e $\Phi(.)$ é uma distribuição de uma variável com distribuição Normal.

A função *hazard* é crescente até atingir um valor máximo a partir do qual se torna decrescente e, $\lim_{t\to 0^+} h(t) = \lim_{t\to\infty} h(t) = 0$.

Tal como as distribuições anteriormente referidas, a distribuição log-normal tem sido frequentemente usada para modelar o tempo de sobrevivência. O maior inconveniente desta distribuição, para situações em que o risco de 'falha' é crescente, deve-se ao facto da função *hazard* ser decrescente a partir de um dado instante; no entanto esta distribuição é adequada quando os valores elevados de t não têm interesse.

2.5.2. Estimação da Função de Verosimilhança num Modelo de Regressão Paramétrico

Em Análise de Sobrevivência os métodos de inferência estatística empregues são, de um modo geral, fundamentados na teoria assintótica da máxima verosimilhança.

No caso de dados completos, para uma determinada família com função densidade de probabilidade f com parâmetros θ (ou vector de parâmetros), a função de verosimilhança $L(\cdot)$ é dada por

$$L(\theta) = \prod_{i=1}^{n} f(t_i;\theta).$$

O estimador de máxima verosimilhança de θ é o valor $\hat{\theta}$ que maximiza a função de verosimilhança $L(\cdot)$. Consideremos ainda função distribuição de tempo de falha T, $F(t_i;\theta)$, sendo θ o parâmetro (ou vector de parâmetros) a estimar.

É importante notar que para outros tipos de dados, que não completos, a expressão $L(\cdot)$ pode não ser definida desta forma. De facto, na construção da função de verosimilhança é preciso ter atenção à existência de censura e/ou truncatura dos dados, pois esta pode alterar a forma da função.

Consideremos o caso de mecanismos de censura do tipo II. Assumindo que $T_1, \dots, T_n$ são independentes e identicamente distribuídas e que seguem uma distribuição contínua com função densidade de probabilidade $f(t)$ e função distribuição $F(t)$, tem-se que a função densidade de probabilidade conjunta de $t_{(1)}, \dots, t_{(r)}$ (estatísticas ordinais) é dada por

$$f_{t_{(1)},\dots,t_{(r)}}((t_1, \dots, t_r); \theta)) = \frac{n!}{(n-r)!}\left[\prod_{i=1}^{r} f(t_i)\right][1 - F(t_r)]^{n-r},$$

com $0 \leq t_1 \leq \cdots \leq t_r$.

Terminada a experiência num instante especifico t_0, no caso de mecanismos com censura do tipo I, a função densidade de probabilidade conjunta é dada por

$$f_{t_{(1)},\dots,t_{(r)}}((t_1, \dots, t_r); \theta)) = \frac{n!}{(n-r)!}\left[\prod_{i=1}^{r} f(t_0)\right][1 - F(t_i)]^{n-r},$$

onde $0 \leq t_1 \leq \cdots \leq t_r \leq t_0$.

2.6. Breve Estudo de Modelos Não Paramétricos

2.6.1. Estimador de Kapan-Meier

A estimação da função de sobrevivência é um aspecto fundamental em estudos de sobrevivência, visto ser utilizada para estimar qual é a probabilidade do evento de interesse não ser observado até um determinado instante. De facto, a estimativa da função de sobrevivência desempenha um papel preponderante na Análise de Sobrevivência, já que é frequentemente usada para sumariar a sobrevivência de determinados grupos de indivíduos. Além disso a sua representação gráfica providencia informação sobre os percentis, a dispersão e, em geral, todos os aspectos da distribuição amostral do tempo de sobrevivência, constituindo uma ferramenta básica para a selecção de um ou mais modelos probabilísticos que possam ser bons ajustes (Andersen & Keiding (2002)).

Perante a inexistência de observações censuradas, numa amostra de dimensão n, a função de sobrevivência, num instante t, pode ser estimada a partir dos tempos de vida observados. Um estimador natural da função de sobrevivência é o estimador empírico, dado por

$$\hat{S}(t) = \frac{número\ de\ observações\ \geq t}{n}, \quad t \geq 0.$$

Porém, perante dados censurados são necessários métodos alternativos. No caso de haver observações censuradas esta estimativa terá de ser alterada porque o número de tempos de sobrevivência maiores ou iguais a t não é, em geral, conhecido.

Kaplan e Meier, em 1958, propuseram um estimador não paramétrico da função de sobrevivência que incorpora censura, designado estimador Kaplan - Meier ou estimador produto - limite, sendo esta uma generalização do estimador empírico para dados censurados.

Sejam $y_1 < y_2 < \cdots < y_k$ os k instantes distintos de eventos de uma amostra de dimensão n $(k \leq n)$, d_i o número de eventos ocorridas em y_i e n_i o número de indivíduos em risco no instante y_i , ou seja, o número de indivíduos 'vivos' e não censurados imediatamente antes do instante y_i, $i = 1, 2, \ldots, k$ (Kaplan & Meier (1958)). O estimador de Kaplan – Meier da função sobrevivência é dada por

$$\hat{S}(t) = \prod_{i:y_i<t} \left(\frac{n_i - d_i}{n_i}\right) = \prod_{i:y_i<t} \left(1 - \frac{d_i}{n_i}\right).$$

A estimativa (curva) $\hat{S}(t)$ é uma função em escada que decresce logo após a observação dos instantes de eventos y_i. Os saltos da função dependem não só do número de eventos, mas também do número de indivíduos em risco no momento, sendo expressos por um factor $1 - \frac{d_i}{n_i}$. A estimativa não decresce nos tempos censurados.

Note-se que se a maior observação registada não for censurada, então $\hat{S}(t)$ toma o valor zero a partir desse instante. No caso em que a maior observação registada seja um tempo de censura, considera-se que $\hat{S}(t)$ está definida apenas até esse instante, nunca atingindo o valor zero.

O estimador de Kaplan – Meier também pode escrever-se recorrendo aos pesos Kaplan – Meier:

$$\hat{S}(t) = 1 - \sum_{i}^{n} W_i I(Y_i \leq t)$$

onde $I(\cdot)$ denota a função indicatriz e W_i os pesos de Kaplan – Meier associados a y_i.

$$\mathrm{W_i} = \frac{\delta_i}{n-i+1}\prod_{j=1}^{i-1} 1 - \frac{\delta_j}{n-j+1}.$$

No caso de não existir censura, este peso é igual a $\frac{1}{n}$.

Por convenção, *os ranks* dos $y_i's$ censurados são maiores que os dos não censurados em caso de empates.

Consideremos hipoteticamente a base de dados, apresentada na Tabela 3.1, apenas com propósito de ilustrar as metodologias referidas do estimador de Kaplan - Meier de acordo com as duas fórmulas alternativas.

Id	***Tempo***	***Falha***	***Censura***
1	4	1	0
2	11	0	1
3	5	1	0
4	1	1	0
5	6	0	1
6	2	1	0
7	2	1	0
8	3	0	1
9	5	0	1
10	2	1	0
11	3	1	0
12	3	1	0
13	6	0	1
14	9	1	0
15	4	1	0

Tabela 2.1- Dados para ilustrar as funções de sobrevivência

Comecemos por ordenar os valores $Y: Y_1 \leq \cdots \leq Y_n$ colocando em caso de empate primeiro os casos não censurados e só depois os censurados.

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
y_i	1	2	2	2	3	3	3^+	4	4	5	5^+	6^+	6^+	9	11^+

Tabela 2.2 - Dados ordenados onde + indica presença de censura pela direita.

Calculando os estimadores para a função de sobrevivência através do método de Kaplan - Meier e do método Kaplan – Meier com pesos obtemos os resultados apresentados na Tabela 2.3.

Identificação (Id)	***Tempo de sobrevivência*** (y_i)	***Número de 'falhas'*** (d_i)	***Número em risco*** (n_i)	***Estatuto*** δ_i	***Pesos K-M*** (W_i)	***Função Distribuição K-M*** $\widehat{F}(t)$	***Função de sobrevivência K-M*** $\widehat{S}(t)$	***Função de sobrevivência K-M com pesos*** $\widehat{S}(t)$
1	1	1	15	1	0.066667	0.066667	0.93333	0.93333
2	2	1	14	1	0.066667	0.133334	0.86666	0.86666
3	2	1	13	1	0.066667	0.200001	0.80000	0.80000
4	2	1	12	1	0.066667	0.266668	0.73333	0.73333
5	3	1	11	1	0.066667	0.333335	0.66667	0.66667
6	3	1	10	1	0.066667	0.400002	0.60000	0.60000
7	3	0	9	0	0	0.400002	0.60000	0.60000
8	4	1	8	1	0.075000	0.475002	0.52500	0.52500
9	4	1	7	1	0.075000	0.550002	0.45000	0.45000
10	5	1	6	1	0.075000	0.625002	0.37500	0.37500
11	5	0	5	0	0	0.625002	0.37500	0.37500
12	6	0	4	0	0	0.625002	0.37500	0.37500
13	6	0	3	0	0	0.625002	0.37500	0.37500
14	9	1	2	1	0.187500	0.812502	0.18750	0.18750
15	11	0	1	0	0	0.812502	0.18750	0.18750

Tabela 2.3- Estimação da função de sobrevivência pelos métodos de Kaplan Meier e Kaplan Meier com pesos

Para além da estimação da função de sobrevivência, em muitas situações interessa também estimar a função de $hazard$ cumulativa $H(t)$,

$$\widehat{H}(t) = -\log \hat{S}(t) = -\sum_{i:y_i \leq t} \log\left(1 - \frac{d_i}{n_i}\right),$$

para $y_i \leq t \leq y_{i+1,}\ i = 1,2,\dots,k$ e $y_1, y_2, \dots, y_k$ são os instantes distintos de ocorrência dos eventos, com $y_{k+1} = \infty$.

2.6.2. Estimador de Kaplan - Meier Modificado

Segundo Bahrawar (2005), em doenças potencialmente fatais, quando uma parte dos dados é censurada o estimador de Kaplan Meier torna-se inviável e ineficiente. Desta forma Bahrawar na sua tese de mestrado apresenta uma modificação ao estimador de Kaplan–Meier de da função sobrevivência. Suponha-se que o conjunto de dados é formado por tempos observados $y_1, y_2, \dots, y_n$, os quais incluem tempos censurados. Ou seja, para alguns dos y_i, é apenas conhecido que para o indivíduo i no tempo y_i ainda não se tinha observado o evento de interesse e desapareceu do estudo depois do tempo y_i . Seja k o número de tempos de censura/falha, e $y_1 < y_2 < \cdots < y_k$ a ordem dos tempos de censura/falha. Defina-se n_i o número de indivíduos ainda 'vivos' imediatamente antes de y_i, incluindo os indivíduos prestes a observar o evento. Defina-se ainda d_i o número de indivíduos que observaram o evento e c_i o número de indivíduos censurados no tempo y_i.

Para cada é atribuído um factor de ponderação w_i dado por tempo y_i

$$w_i = \left\{1 - \frac{c_i}{n_i}\right\}$$

sendo conhecido como taxa de não censura. Onde $w_i = 1$ se não há tempo de censura no instante y_i (pois $c_i = 0$), e $w_i < 1$, no caso de censura no tempo y_i.

O estimador de Kaplan –Meier Modificado proposto por Bahrawar (2005), é definido por

$$\widehat{\boldsymbol{S}}(\boldsymbol{t}) = \begin{cases} \mathbf{1} & \boldsymbol{se}\ \ \boldsymbol{t} = \mathbf{0} \\ \prod_{\boldsymbol{i}:\boldsymbol{y_i}<\boldsymbol{t}} \left(\mathbf{1} - \frac{\boldsymbol{c_i}}{\boldsymbol{n_i}}\right)\left(\mathbf{1} - \frac{\boldsymbol{d_i}}{\boldsymbol{n_i}}\right). \\ \mathbf{0} & \boldsymbol{se}\ \boldsymbol{t} \geq \boldsymbol{y_k} \end{cases}$$

Neste caso, o estimador de Kaplan – Meier modificado da função de sobrevivência fica assim definida para todo o $t \geq 0$, alcançando o valor zero, mesmo no caso em que a última estimação seja censurada. Baharawar (2005) sugere que se considere a última observação censurada como tempo de falha, para garantir a sua definição em todos os instantes. Se não há censura, $\hat{S}(t)$ coincide com a função empírica de sobrevivência, como $W_i = 1$ e $\hat{S}(t)$ coincide com Kaplan – Meier, que por sua vez se converte na função de sobrevivência empírica se todos os dados são completos (Huang *et al.* (2000)).

Consideremos novamente os dados da Tabela 2.1 de forma a ilustrar as metodologias referidas do estimador de Kaplan - Meier e Kaplan - Meier modificado.

Organizando os dados da Tabela 2.1 e calculando os estimadores para a função de sobrevivência através do método de Kaplan – Meier e do método Kaplan – Meier modificado obtemos os resultados presentes na Tabela 2.4.

Comparando os resultados do estimador de Kaplan – Meier e Kaplan Meier modificado, Tabela 2.4, podemos constatar que com o estimador de Kaplan – Meier se obtém uma maior probabilidade de sobrevivência. De notar, ainda, que no estimador de Kaplan – Meier modificado a probabilidade de sobrevivência da última observação censurada é zero.

No entanto, se os tempos (de evento e de censura) forem todos distintos, sem empates, o estimador que se obtém é o estimador empírico.

Número i	***Tempo de sobrevivência*** (t_i)	***Número de mortes*** (d_i)	***Número de censuras*** (c_i)	***Número em risco*** (n_i)	***Pesos das observações censuradas*** (W_i)	***Função de sobrevivência K-M*** $\widehat{S}(t)$	***Função de sobrevivência K-M modificada*** $\dot{S}(t)$
1	1	1	0	15	1.00000	0.93333	0.93333
2	2	3	0	14	1.00000	0.73333	0.73333
3	3	2	1	11	0.90909	0.59999	0.54545
4	4	2	0	8	1.00000	0.44999	0.37125
5	5	1	1	6	1.00000	0.37499	0.28125
6	6	0	2	4	0.50000	0.37499	0.12784
7	9	1	0	2	1.00000	0.18750	0.02905
8	11	0	1	1	0.00000	0.18750	0.00000

Tabela 2.4 - Estimação da função de sobrevivência pelos métodos de Kaplan Meier e Kaplan Meier modificado

2.6.3. Estimação de $\widehat{Var}\left[\widehat{S}(t)\right]$

Breslow e Crowley (1974) e Meier (1975) provaram que sobre condições muito gerais, o estimador $\hat{S}(t)$ pode ser considerado como um estimador de máxima verosimilhança não paramétrica de $S(t)$. A partir da metodologia de máxima verosimilhança chega-se à seguinte expressão para a estimativa da variância do estimador $\hat{S}(t)$

$$\widehat{Var}\left[\hat{S}(t)\right] = \left[\hat{S}(t)\right]^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)},$$

conhecida como a formula de Greenwood's (Greenwood, 1926).

Tsiatis (1981), sugere sobre condições assintóticas o denominador $n_i{}^2$ em vez de $n_i(n_i - d_i)$. Fleming e Harrington (1991) e Andersen (1993) apresentam rigorosas derivações desta fórmula bem como cálculos de viés e teoremas limite que justificam a normalidade assintótica. Klein (1991) discutiu o comportamento do viés em pequenas amostras dos estimadores de variância. As suas conclusões indicam que a fórmula de Greenwood é preferível, e é suficientemente exacta desde que o tamanho esperado do conjunto de risco seja pelo menos cinco (Vernables & Ripley (1999)).

2.6.4. Intervalo de Confiança para $\widehat{S}(t)$

Um intervalo de confiança calculado num determinado instante, pertencente ao domínio da curva de sobrevivência, indica a credibilidade da estimativa nesse instante. Esse intervalo pode ser calculado no ponto de interesse considerando a forma habitual dos intervalos de confiança e assumindo que o estimador de Kaplan – Meier tem distribuição assintótica normal. Isto significa que o intervalo com $(1 - \alpha) \times 100\%$ de confiança para a função de sobrevivência num instante específico é

$$\hat{S}(t) \pm \mathrm{Z}_{\frac{\alpha}{2}} \sqrt{\widehat{Var}_G\left[\hat{S}(t)\right]}$$

onde $\mathrm{Z}_{\frac{\alpha}{2}}$ é o quantil de probabilidade $\frac{\alpha}{2}$ de uma distribuição normal reduzida. A estimativa de variância de $\hat{S}(t)$ aqui sugerida é a proposta por Greenwood (1926).

Os intervalos de confiança obtidos por este processo são simétricos, sendo possível que os limites do intervalo de confiança sejam maiores do que um ou menores do que zero, nos instantes em que $\hat{S}(t)$ está próximo desses valores. Uma solução pragmática para este problema, é a substituição de qualquer limite maior que um por 1 e qualquer limite menor que zero por 0.

Uma alternativa é transformar $S\,(t)$ num valor pertencente ao intervalo $(-\infty, +\infty)$, e obter um intervalo de confiança para o valor transformado. Existem várias transformações possíveis, sendo três delas a transformação log , $log(-log)$ e linear.

Transformação	
linear	$\hat{S}(t)\left[1 \pm \mathrm{Z}_{\frac{\alpha}{2}}\hat{\sigma}(\widehat{H}(t))\right]$
log	$\hat{S}(t)\exp\left[1 \pm \mathrm{Z}_{\frac{\alpha}{2}}\hat{\sigma}(\widehat{H}(t))\right]$
$log(-log)$	$\exp\left\{-\widehat{H}(t)\exp\left[\pm \mathrm{Z}_{\frac{\alpha}{2}}\frac{\hat{\sigma}(\widehat{H}(t))}{\widehat{H}(t)}\right]\right\}$

Tabela 2.5 - Transformações da função de sobrevivência

2.6.5. Estimador de Kaplan – Meier com Estratificação

Em Análise de Sobrevivência interessa analisar os factores endógenos ou exógenos aos indivíduos que contribuem para a ocorrência do acontecimento de interesse, isto é, características como o sexo, a idade, a utilização de determinado fármaco, entre outras, podem ter um papel primordial no tempo de sobrevivência, e originar curvas de sobrevivência distintas. A estratégia utilizada, com base no estimador de Kaplan-Meier, para comparar as diferentes curvas correspondentes aos vários grupos, é a estratificação. Esta estratificação consiste na divisão do conjunto total de observações em grupos distintos, de acordo com as covariáveis de interesse, e na estimação das funções de sobrevivência, separadamente para cada um dos grupos, o que possibilita, de modo informal, avaliar se a variável tem influência no tempo de sobrevivência. O estimador Kaplan – Meier é assaz vantajoso, na medida em que possibilita interpretações e representações gráficas intuitivas.

Depois das funções de sobrevivência estimadas e representadas graficamente será importante testar se existem entre as curvas de sobrevivências diferenças significativas desses diversos grupos.

Para comparar as funções de sobrevivência são geralmente utilizados testes não paramétricos, designados por teste de ordem (*rank test*), visto a estatística de teste depender unicamente das ordens das observações.

Deste modo, avaliam se existem diferenças significativas nos tempos de sobrevivência entre dois ou mais grupos diferentes. Dos testes não paramétricos existentes optou-se por utilizar o

teste *log–rank* dado estar disponível no *software* R. O teste referido é usados para testar a igualdade de funções de sobrevivência, ou seja, a hipótese nula a testar é

$$H_0: S_1(t) = S_2(t) = \cdots = S_i(t), \text{ onde } i \in \{2,3, \dots\}.$$

2.6.6. Teste *Log – Rank*

Baseado no trabalho de Mantel e Haenzel (1959) foi proposto o teste, geralmente, designado por teste *log – rank* ou teste de Mantel – Haenszel.

Suponha-se que se tem dois grupos (1 e 2) a que correspondem duas amostras de *m* e *n* indivíduos, respectivamente, e que $t_1 < t_2 < \cdots < t_k$ são os k instantes dos evento relativos aos $m + n$ indivíduos. Sabe-se que existem $n_{i,j}$ indivíduos em risco no grupo i, $(i = 1,2)$, imediatamente antes do instante t_j, e que nesse instante observaram o evento d_{1j} indivíduos do grupo 1 e d_{2j} indivíduos do grupo 2, para $j = 1, 2, \dots, k$. Assim, $d_j = d_{1j} + d_{2j}$ corresponde ao número total de mortes ocorridas nos dois grupos, no total $n_j = n_{1j} + n_{2j}$ indivíduos em risco, no instante t_j.

A informação relevante em cada instante t_j $(j = 1,2, \dots, k)$ pode ser resumida numa tabela de contingência 2×2:

Grupo	**Nº de mortes em t_j**	**Nº de sobreviventes para além de t_j**	**Nº de indivíduos em risco em t_j**
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Tabela 2.6 - Tabela de contingência 2×2

Considera-se a hipótese nula, de igualdade das funções de sobrevivência nos dois grupos distintos. Então, supondo H_0 verdadeira, a distribuição de d_{1j}, condicional aos totais marginais, é hipergeométrica, sendo que

$$p(d_{1j}|d_j, n_j) = \frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}$$

e, por conseguinte sob H_0, o valor médio condicional da variável aleatória d_{1j} é dado por $e_{ij} = \frac{n_{1j}d_j}{n_j}$, que representa o número esperado de indivíduos para quais se observa o evento no instante t_j no grupo 1. A variância condicional de d_{1j} é dada por

$$v_{1j} = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}.$$

Com o objectivo de obter uma medida global do desvio dos valores observados de d_{1j} em relação aos valores esperados é importante combinar a informação das k tabelas de contigência 2×2, considerando a estatística

$$U_L = \sum_{j=1}^{k} (d_{1j} - e_{1j}),$$

em que $\sum d_{1j} - \sum e_{1j}$ é a diferença entre o total eventos observados e esperados no grupo 1. Dado que se pode assumir a independência entre os eventos ocorridas nos k instantes, a variância de U_L é apenas a soma das variâncias de d_{1j}, sendo assim

$$Var(U_L) = \sum_{j=1}^{k} v_{1j} = V_L.$$

A estatística de teste proposta por Mantel e Haenszel (1959) é dada por

$$C_L = \frac{U_L^2}{V_L},$$

que sobe a hipótese de H_0 tem distribuição assintótica Qui - quadrado com 1 grau de liberdade χ_1^2.

A hipótese H_0 é rejeitada para valores grandes da estatística, ou seja, para $W_L > \chi_{1-\alpha}^2$, sendo $\chi_{1-\alpha}^2$ o quantil de probabilidade $1 - \alpha$ da distribuição Qui - quadrado com 1 grau de liberdade.

2.6.7. Estimador de Nelson-Aalen e Fleming-Harrington

O estimador proposto por Kaplan e Meier (1958) é o estimador mais utilizado para a estimação não paramétrica da função de sobrevivência. Frequentemente, interessa também estimar a função de risco cumulativa. Um estimador natural de H(t) será $\widehat{H}(t) = -log\,\hat{S}(t)$, onde $\hat{S}(t)$ é o estimador de Kaplan – Meier.

Um estimador alternativo é o sugerido por Nelson (1972) e estudado por Aalen (1978) torna-se cada vez mais usual. Este estimador é designado por estimador de Nelson-Aalen.

Suponhamos que $t_1, t_2, \dots, t_k$ são os k instantes de morte distintos de uma amostra de dimensão n $(k \leq n)$ tal que $t_1 < t_2 < \cdots < t_k$. Seja d_i o número de mortes ocorridas em t_i e n_i o número de indivíduos em risco no instante $t_i, i = 1,2, \dots, k$. O estimador de Nelson-Aalen da função risco cumulativa é dado por

$$\widehat{H}(t) = \sum_{i:t_i \leq t} \frac{d_i}{n_i},$$

e a estimativa da variância de $\widehat{H}(t)$ é

$$Var\left[\widehat{H}(t)\right] = \sum_{t_i \leq t} \frac{d_i}{n_i}.$$

O estimador de Nelson-Aalen estima directamente a função de risco cumulativa, embora como é óbvio se possa obter também uma estimativa da função de sobrevivência. Assim sendo, o estimador de Nelson-Aalen para a função de sobrevivência, também conhecido por estimador de Breslow, é dado por

$$\hat{s}_{NA}(t) = e^{\sum_{i:t_i \leq t} \left(\frac{d_i}{n_i}\right)}$$

Embora o estimador de Nelson-Aalen apresente um melhor comportamento, para pequenas amostras, do que o estimador de Kaplan-Meier, em muitas circunstâncias as estimativas serão muito semelhantes, principalmente quando ainda existem muitos indivíduos em risco.

Baseado no estimador Nelson - Aalen, chegamos a um estimador alternativo ao estimador de Kaplan – Meier, o estimador de Fleming – Harrington para a função de sobrevivência, dado por $\hat{S}(t) = e^{-\widehat{H}(t)}$.

A estimativa de sobrevivência de Fleming – Harrington (1991) foi proposta para colmatar o problema da estimativa de Nelson (1972) ser susceptível à ocorrência de empates, podendo perante um grande número de empates levar a um enviesamento significativo.

2.7. Covariáveis

O objectivo fundamental da Análise de Sobrevivência é a análise do tempo de sobrevivência de cada indivíduo. É razoável admitir que esse tempo depende de inúmeros factores, tais como tratamentos, características dos indivíduos (sexo, idade, etc.) e variáveis exógenas.
A partir da década de 70, foi dada uma ênfase particular ao estudo da associação entre o tempo de sobrevivência e variáveis designadas por explicativas ou covariáveis, representando os factores acima mencionados. Sempre que tal for possível, os valores individuais destas variáveis devem ser tomados em consideração, visto fornecerem informação acerca da heterogeneidade existente na população.
As covariáveis podem ser dependentes do tempo ou constantes (fixas ao longo do tempo), consoante os seus valores são medidos ao longo do período de observação ou uma única vez. As variáveis explicativas podem ser classificadas como covariáveis externas ou internas. Uma análise mais detalhada sobre a classificação de covariáveis será realizada posteriormente.

2.8. Modelo de Regressão de Cox

Os estudos em análise de sobrevivência envolvem muitas vezes covariáveis que podem estar relacionadas com o tempo de sobrevivência. Essas covariáveis devem ser incluídas na análise estatística dos dados para explicar o seu possível efeito no tempo de sobrevivência. Uma das alternativas metodológicas que incorpora informações no estudo do tempo de sobrevivência através da introdução de covariáveis é o modelo de riscos proporcionais. De um modo geral, este problema pode ser abordado através de um estudo de um modelo de regressão que tenha em consideração a ocorrência de observações censuradas.

Uma escolha comum para fazer esta relação é dada pelo modelo hazard proporcional, introduzido por Cox (1972). A importância do modelo de Cox deve-se ao facto deste ser modelo semi-paramétrico, de fácil e simples interpretação e no qual a ocorrência de censura é facilmente acomodada. Para além disso encontra-se disponível na maioria dos programas estatísticos. No entanto, apesar destas vantagens, não deixa de ser necessário algum cuidado na sua utilização. Assim ao longo desta secção, não só fazemos uma descrição deste modelo, como também apresentamos alguns métodos gráficos e testes de hipóteses para a sua validação.

O modelo de Cox assume que a função *hazard* se pode escrever na forma,

$$h(t;z) = h_0(t)e^{\beta^T Z},$$

onde $h_0(t)$ é uma função *hazard* suporte arbitrária não negativa, β um vector de coeficientes de regressão e Z um vector de covariáveis.

Note-se que a razão das funções *hazard* para dois indivíduos com vectores de covariáveis fixos é constante no tempo - é este o motivo pelo qual o modelo é vulgarmente conhecido por modelo de *hazards* proporcionais. Esta razão não depende do tempo, isto é, o risco de falha de um indivíduo em relação ao outro é constante para todos os tempos de acompanhamento. Os dois componentes multiplicativos do modelo são de naturezas distintas, um não-paramétrico e o outro paramétrico sendo esta a razão do modelo ser do tipo semi-paramétrico o que o torna bastante flexível.

Assim duas características deste modelo são:

- as funções de risco correspondentes a dois indivíduos diferentes com vectores de covariáveis Z_1 e Z_2 são proporcionais, pelo que a razão entre estas funções em qualquer instante t

$$\frac{h(t;z_1)}{h(t;z_2)} = \frac{h_0(t)e^{\beta^T Z_1}}{h_0(t)e^{\beta^T Z_2}} = e^{[\beta^T (Z_1 - Z_2)]}$$

não depender de t;

- as covariáveis afectam a função hazard de modo multiplicativo de acordo com o factor $e^{\beta^T Z}$, que é designado por risco relativo.

O modelo de regressão de Cox é caracterizado pelos coeficientes β que medem o efeito das covariáveis sobre a função de risco. Dessa maneira é necessário um método de estimação para se fazer inferência no modelo. Cox (1972) construiu uma função de verosimilhança que não

depende de $h_0(t)$, permitindo assim a realização de inferências sobre β sem que seja necessário especificar $h_0(t)$.

Considerando n indivíduos em estudo e k instantes distintos de observação de eventos, tal que $t_1 < t_2 < \cdots < t_k$, em que $k \leq n$, o conjunto de risco no instante t_i, designado por R_i é definido por

$$R_i = R(t_i) = \{j: t_j \geq t_i\},$$

sendo o conjunto de índices associados aos indivíduos em observação imediatamente antes do instante t_i.

A função de verosimilhança, proposta por Cox (1972) para a realização de inferência sobre β, é dada por

$$L(\beta) = \prod_{i=1}^{k} \frac{e^{(\beta^T Z_i)}}{\sum_{l \in R_i} e^{(\beta^T Z_l)}},$$

em que z_i é o vector de variáveis explicativas associado ao indivíduo com evento no instante t_i. Cox (1975) designou a função $L(\beta)$ por função de verosimilhança parcial. Assim, mostrou que, embora não se trate de uma função no sentido usual, dado não permitir a obtenção de um estimador do vector de parâmetros β e, sob vertas condições bastante gerais, verifica as propriedades usuais dos estimadores de máxima verosimilhança.

Note-se que a função de verosimilhança proposta por Cox (1984) não depende de $h_0(t)$, o que permite a realização de inferência sobre o vector de parâmetros β, sem que seja necessário fazer restrições à forma de $h_0(t)$.

Este facto, faz com que o método de máxima verosimilhança usual não possa ser utilizado, pois a presença da componente não-paramétrica $h_0(t)$ na função de verosimilhança torna este método inapropriado. Frente a tal dificuldade, Cox (1975) propôs o método de máxima verosimilhança parcial que condiciona à verosimilhança a história dos tempos de sobrevivência e censuras anteriores e desta forma elimina a função de base desconhecida (Ferreira (2007)). Tal como Cox (1975) mostrou, o método de construção de verosimilhança parcial, sob condições de regularidade bastante fracas, leva às propriedades assintóticas usuais de inferência baseada na verosimilhança.

Na situação em que tenha ocorrido falha de mais do que um indivíduo em simultâneo dando origem a valores iguais, a função de verosimilhança parcial não poderá ser aplicada. Nesta situação, para os n indivíduos em estudo, suponhamos que foram observados eventos nos instantes $t_1 < t_2 < \cdots < t_k$. Seja d_i o número eventos ocorridos no instante t_i, e Z_{ij} o vector de covariáveis explicativas associadas ao indivíduo j, que cujo evento ocorre em t_i, $j =$

$1, \dots, d_i,\ i = 1, \dots, k$. Se, quando comparando com o número de indivíduos pertencentes ao conjunto de risco $R_i,\ i = 1, \dots, k,$ o número de d_i de indivíduos com evento em t_i for pequeno, então a função de verosimilhança parcial pode ser aproximada pela função, proposta por Breslow (1974),

$$L(\beta) = \prod_{i=1}^{k} \frac{e^{(\beta^T s_i)}}{\left[\sum_{l \in R_i} e^{(\beta^T Z_l)}\right]^{d_i}}.$$

Onde $s_i = \sum_{j=1}^{d_i} Z_{ij}$, para $i = 1, \dots, k$. Esta função de verosimilhança é habitualmente usada no *software* estatístico. Caso sejam observados um número substancial de tempos de sobrevivência iguais, então será aconselhável optar por um modelo discreto.

Uma desvantagem do modelo de Cox é a sua vulnerabilidade a certos problemas de consistência. Revela-se assim da maior importância o desenvolvimento de modelos alternativos a serem aplicados em situações em que as hipóteses do modelo de Cox não são válidas. De facto, o modelo apresentado é válido quando se considera um vector de p covariáveis, no entanto a proporcionalidade entre as funções *hazard* correspondentes a indivíduos diferentes não se mantém, em geral, quando é omissa alguma covariável, ainda que esta seja independente das restantes. O mesmo problema se põe quando se verifica alterações na precisão com que são medidas as covariáveis (Segão, (2000)). Não convém esquecer que o modelo de *hazards* proporcionais de Cox é condicional ao valor observado do vector das covariáveis. Portanto, se uma covariável é medida com erro ou é mal classificada, o estimador de verosimilhança parcial não será a melhor opção, visto que conduzirá a resultados enviesados (Struthers & Kalbfleisch (1986)).

2.9. Covariáveis Dependentes no Tempo

Os valores das covariáveis são medidos durante o período de observação de forma a modelar dados de sobrevivência. Contudo, existem covariáveis que são observadas durante o estudo e cujos valores mudam nesse período. Surgem assim as covariáveis dependentes no tempo que se caracterizam por variáveis explicativas cujos valores podem mudar ao longo do tempo.

É possível distinguir dois tipos de covariáveis dependentes no tempo, que de acordo com Kalbfleisch e Prentice (1980) são identificadas como externas ou internas. Entende-se por

covariáveis externas as covariáveis que não estão envolvidas em processos de falha, não requerendo necessariamente a sobrevivência do paciente para a sua existência. As covariáveis internas são aquelas onde a mudança da covariável ao longo do tempo está relacionada com a sobrevivência do indivíduo, os valores observados levam informação sobre o seu tempo de sobrevivência. Só podem ser medidas enquanto o paciente sobrevive.

As covariáveis dependentes no tempo podem ser empregues tanto para acomodar medidas que variam com o tempo durante o estudo, como também podem ser relevantes para modelar o efeito de indivíduos que mudam de grupo durante um tratamento.

2.10. Modelo de Cox com Covariáveis Dependentes no Tempo

O modelo de regressão de Cox pode ser generalizado de forma a incorporar covariáveis dependentes no tempo.

$$h(t;z) = h_0(t)e^{\beta^T Z(t)}$$

onde $Z(t)$ é o valor da covariável no tempo t.

Com a incorporação de covariáveis dependente no tempo ao modelo de regressão de Cox, o modelo apresentado já não verifica a condição de risco proporcional. Os valores dos vectores de covariáveis Z_1 e Z_2 dependem do tempo t e a razão das funções de risco no tempo t para dois indivíduos diferentes dada por

$$\frac{h(t;Z_1)}{h(t;Z_2)} = e^{[\beta^T(Z(t)-Z_2(t))]}$$

é dependente no tempo.

As estimativas dos parâmetros do modelo de regressão de Cox com covariáveis dependentes no tempo podem ser obtidas generalizando a função de verosimilhança parcial para

$$L(\beta) = \prod_{i=1}^{k} \frac{e^{(\beta^T Z_i(t_i))}}{\sum_{l\in R_i} e^{(\beta^T Z_l(t_i))}}.$$

Uma boa aproximação para o logaritmo da função verosimilhança parcial, quando os dados não apresentam muitos empates é dada por

$$\sum_{i=1}^{k} \beta^T s_i(t_i) - d_i \log \sum_{j \epsilon \mathcal{R}(t_i)} e^{\beta^T Z_j(t_i)} .$$

2.11. Breve Referência aos Processos de Escolha do Modelo de Regressão de Cox

A escolha do melhor modelo de regressão de Cox é uma difícil tarefa pois raramente existe um único melhor modelo. Por conseguinte, é necessário estabelecer uma estratégia que seja um compromisso entre a eleição de um modelo que proporcione um bom ajuste aos dados e, por outro lado, seja fácil de interpretar suavizando os dados. Para tal é habitual utilizar os procedimentos de busca sequencial.

Os métodos de busca sequencial têm em comum a abordagem geral de estimar a equação de regressão com um conjunto de variáveis e então acrescentar selectivamente ou eliminar variáveis até que alguma medida de critério geral seja alcançada. Esta abordagem fornece um método objectivo para seleccionar variáveis maximizando a previsão com o menor número de variáveis empregues (Hair *et al.* (1998)). A abordagem sequencial mais comum para a selecção de variáveis é a estimação *stepwise*. Os procedimentos stepwise dividem-se em *forward selection*, método composto e *backward elimination.* Para mais detalhes consultar Mason *et al.* (2003) .

O critério de informação de *Akaike* é outro possível critério de selecção de um modelo. Proposto por *Akaike* (1974), baseia-se na função log-verosimilhança com a introdução de um factor de correcção como modo de penalização da complexidade do modelo, examina a seguinte estatística

$$AIC = -2loglikelihood + 2 \times r$$

onde r é o número de parâmetros do modelo. Um valor baixo para AIC é considerado como representativo de um melhor ajustamento e, assim, na selecção de modelos deve-se ter como objectivo a minimização de AIC.

Existem vários métodos para verificar a validade do pressuposto de *hazards* proporcionais. Esse pressuposto pode ser validado através de um gráfico $\log(-\log(S(t)))$ versus t. As curvas obtidas devem ser paralelas. Assim, o modelo de hazards proporcionais é inadequado quando as curvas se intersectam. Outra aproximação que pode ser utilizada é o gráfico dos resíduos de Schoenfeld. Estes resíduos, propostos por Schoenfeld (1982), são muito úteis na avaliação da hipótese de riscos proporcionais, após o ajustamento de um modelo de Cox. A cada indivíduo não corresponde apenas um resíduo mas um conjunto de valores, onde cada valor é referente a cada uma das variáveis explicativas incluídas no modelo de regressão de Cox. Também apresenta a vantagem de não ser necessário obter uma estimativa da função de risco cumulativa.

Capítulo 3

Modelos Multiestado

3.1. Generalidades sobre estes modelos

Os modelos multiestado podem ser considerados como uma generalização do modelo de sobrevivência. Enquanto um modelo de sobrevivência considera apenas o tempo para determinado evento, os modelos multiestado lidam com sistemas de eventos relacionados. Estes sistemas de eventos são descritos como definindo um conjunto de estados e visualizando as transições entre os estados. Neste contexto, os modelos multiestado são, frequentemente utilizados para descrever dados longitudinais e definidos a partir de um processo estocástico em tempo contínuo permitindo que os indivíduos se movam através de um número finito de estados. A complexidade de um modelo multiestado depende do número de estados definidos e do número de transições permitidas entre eles (Meira-Machado *al et.* (2009)). Exemplos comuns de estados são condições como ‘saudável’, ‘doente’, ‘doente com complicação’ e ‘morto’.

Os estados podem ser considerados como transientes ou absorventes (Meira-Machado (2006)). Um estado absorvente não permite transições a partir dele sendo que quando um indivíduo atinge esse estado permanece nele para sempre. A morte é o exemplo mais comum de um estado absorvente. Um estado diz-se transiente quando absorvente possível ocorrer transições a partir dele..

A experiência de um indivíduo pode ser encarada como um processo que envolve dois estados, com uma única transição possível de 'vivo' a 'morto'. No entanto, o estado que caracteriza os indivíduos 'vivos' pode ser particionado em dois ou mais estados intermédios. Nestes modelos usualmente a 'morte' é o evento de interesse, mas outros estados intermédios são identificados. Por exemplo, em estudos de diversos cancros, pode ser observados estados intermédios tais como; 'vivo com recorrência', 'vivo com metástases', etc.

Em geral, um evento pode ser representado por uma transição de um estado para o outro e, quando são observados vários eventos, os modelos multiestado permitem uma modelação apropriada.

Os modelos multiestado facultam uma melhor compreensão do progresso do evento de interesse proporcionando uma utilização mais eficiente da informação incompleta, quando porções da história da doença de um indivíduo são conhecidas. A utilização de covariáveis nas intensidades de transição podem também explicar diferenças nesse progresso entre a população, bem como revelar que diferentes covariáveis afectam diferentes transições.

Nas secções subsequentes, abordam-se os modelos multiestado usando terminologia clínica. No entanto, salienta-se que esta abordagem se mantém válida se aplicada a eventos de interesse fora desse âmbito.

3.2. Formulação do Modelo Multiestado

Um modelo multiestado é um processo estocástico $(X(t), t \in [0, \infty))$ em que $X(t)$ denota uma variável aleatória que pode assumir um dos valores no conjunto finito de estados $S = \{1, \dots, N\}$. Este processo possui a informação das diferentes transições que ocorrem a um indivíduo ao longo do tempo, assim como o tempo de transição.

Os processos multiestado são completamente caracterizados por intensidades de transição ou por probabilidades de transição entre estados h e j, $(h, j \in S)$. Para $s \leq t$, definimos as probabilidades de transição entre estados h e j por

$$p_{hj}(s,t) = p(X(t) = j | X(s) = h, \mathcal{H}_{s^-})$$

para $s, j \in [0, \infty)$. Com a evolução do processo ao longo do tempo, uma história $\mathcal{H}_{s^-}$ (uma $\sigma - algebra$) do processo é gerada no instante imediatamente anterior a s, representada por s^-, consistindo na observação do processo no intervalo $[0, s)$.

O risco instantâneo de progressão do estado h para o estado j, representado pelas funções intensidade de transição, é dado por

$$\alpha_{h,j}(t) = \lim_{\Delta t \to 0} \frac{p_{hj}(t, t + \Delta t)}{\Delta t}.$$

Os modelos multiestado, são baseados em processos de contagem multivariados, e adaptam-se a modelações mais complexas de eventos históricos. Diferentes suposições podem ser feitas acerca das intensidades de transição.

Os modelos de multiestado não são equivalentes aos modelos de Markov, no entanto ambos partilham o conceito de estado. A suposição mais simples do modelo de Markov é que a evolução futura depende apenas do estado ocupado no tempo actual t. Por outras palavras, o passado do processo, ou história, é resumido pelo estado no tempo t. Estes modelos não satisfazem obrigatoriamente estes pressupostos. Na realidade, outros factores podem desempenhar um papel na evolução do processo, mas o estado no tempo t ainda é a principal informação a ter em conta (Commenges (1999)). Podemos diferenciar os modelos homogéneos no tempo, os modelos de Markov e os modelos semi-Markov para modelos de multiestado. Nos modelos homogéneos no tempo as intensidades de transição não dependem do tempo, sendo constantes ao longo do tempo, ou seja, independente de t. Num modelo multiestado de Markov assume-se que a intensidade de transição apenas depende da história do processo pelo estado ocupado no tempo *t*, ou seja é independente da história do processo. Num modelo multiestado semi-Markov a evolução futura não depende apenas do estado actualmente ocupado, h, mas também do tempo de entrada no estado actual, t_h . Nestes casos representamos as intensidades de transição por $\alpha_{hj}(t, t - t_h)$.

3.3. Representação de Modelos Multiestado

Os modelos multiestado são representados por diagramas que indicam o número finito de estados clínicos que um paciente pode ocupar. De seguida ilustram-se alguns desses modelos. Uma análise mais detalhada, destes e outros modelos pode ser encontrada em Hougaard (1999).

3.3.1. Modelo de Mortalidade

O modelo de mortalidade é o caso mais simples de entre os modelos multiestado e ocorre quando existe apenas uma possibilidade de transição entre os estados inicial 'vivo' e um estado absorvente 'morto' (Andersen & Keiding (2002)). Tal como sugere, este modelo é caracterizado por apenas dois estados.

Figura 3.1- Esquema do modelo de mortalidade

3.3.2. Modelo de três Estados Progressivos

O modelo de três estados progressivos é apresentado na Figura 3.2. Particionando o estado inicial 'vivo' em dois estados transientes surge o modelo de três estados progressivos. Cada indivíduo em estudo só poderá transitar para o estado mais avançado subsequente.

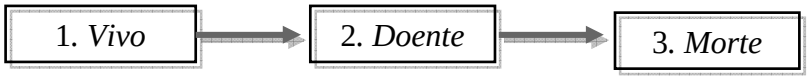


Figura 3.2 - Esquema modelo 3 estados progressivos

3.3.3. Modelo illness-death

O modelo enfermidade-morte, '*illness-death*'. A sua representação é dada na Figura 3.3.

Neste modelo, cada indivíduo pode transitar entre os estados de 'saudável' e 'doente' até entrar no estado absorvente de 'morte'. Como se pode constatar na Figura 3.3, está presente a possibilidade de reversibilidade do modelo, ou seja, é possível ocorrer a transição a partir de 2 para 1. No modelo *illness-death* progressivo, a transição de 2 para 1 não é possível.

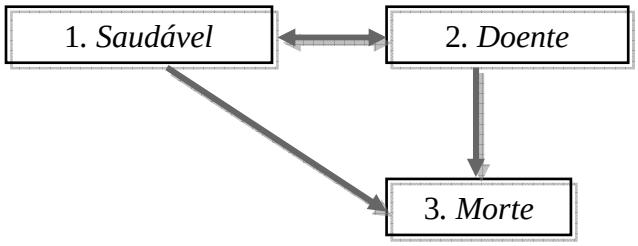


Figura 3.3 - Esquema do modelo *illness-death*

3.3.4. Modelo Bivariado

O modelo bivariado é um modelo multiestado para dados paralelos bivariados. Este modelo pode ser utilizado para, por exemplo, estudar a sobrevivência de gémeos onde são observados quatro estados: ambos 'vivos', o indivíduo 1 é o único que verifica o evento 'morte', o indivíduo 2 é o único que evidencia o evento 'morte', e ambos 'mortos'.

Figura 3.4 - Esquema do modelo bivariado

3.4. Breve Revisão sobre Probabilidades de Transição

3.4.1. Modelo de Markov

O modelo de Markov é o mais popular, sendo muitas vezes aplicado devido à sua simplicidade.

O processo $(X(t), t \geq 0)$ é um processo de Markov, entre os estados h e j, se para qualquer s, t ($0 \leq s < t$) temos

$$p_{hj}(s,t) = P(X(t) = j | X(s) = h, \mathcal{H}_{s^-}) = P(X(t) = j | X(s) = h),$$

desta forma, o futuro do processo depois do tempo s depende apenas do estado ocupado no tempo s.

É importante estimar as probabilidades de transição dado que estas fornecem informações que possibilitam fazer previsões a longo prazo. Num modelo de Markov, a probabilidade de transição pode ser calculada a partir das intensidades de transição resolvendo a equação diferencial intitulada por *forward Kolmogorov equation* (Meira-Machado *al et.* (2009)).

Segundo Hougaard (1999), a função hazard da transição do estado h para o estado j é definida por

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{P(X(t + \Delta t) = j | X(s) = h)}{\Delta t}$$

para qualquer s, t ($0 \leq s < t$).

A título de exemplo, vamos aplique-se o descrito ao modelo *illness-death* progressivo, um modelo com os três estados 'saudável', 'doente' e 'morte'. Para o tempo t, definam-se os estados

$$S(t) = \begin{cases} 1 & \text{se o indivíduo está no estado } 'saudável' \\ 2 & \text{se o indivíduo está no estado } 'doente' \\ 3 & \text{se o indivíduo está no estado } 'morte' \end{cases}$$

e as intensidades de transição

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{P(X(t + \Delta t) = j | X(s) = h)}{\Delta t}$$

onde $h = 1,2$, $j = 1,2,3$ e $h \leq j$.

Dado o estatuto de um indivíduo, em qualquer tempo t, podemos obter a probabilidade deste estar em qualquer estado elegível no tempo $s > t$. Denotando as probabilidades de transição com $p_{hj}(s,t) = P(X(t) = j | X(s) = h)$, a matriz de probabilidade de transição é definida por

$$\begin{bmatrix} p_{11}(s,t) & p_{12}(s,t) & p_{13}(s,t) \\ 0 & p_{22}(s,t) & p_{23}(s,t) \\ 0 & 0 & 1 \end{bmatrix}$$

Nos modelos de Markov, as probabilidades de transição estão relacionados com as intensidades de transição através das equações diferenciais de *Kolmogorov*. Para o modelo *illness-death* progressivo as equações de *Kolmogorov* (Hougaard (1999)), são dadas por:

$$\frac{dp_{11}}{dt}(s,t) = -p_{11}(s,t)[\alpha_{12}(t) + \alpha_{13}(t)],$$

$$\frac{dp_{22}}{dt}(s,t) = -p_{11}(s,t)\alpha_{12}(t),$$

$$\frac{dp_{12}}{dt}(s,t) = p_{11}(s,t)\alpha_{12}(t) - p_{12}(s,t)\alpha_{12}(t).$$

A solução para estas equações é

$$p_{11}(s,t) = e^{-(A_{12}(s,t)+A_{13}(s,t))}$$

$$p_{22}(s,t) = e^{-(A_{23}(s,t))}$$

$$p_{12}(s,t) = \int_s^t p_{11}(s,u)\alpha_{12}(u)p_{22}(u,t)du$$

onde $A_{hj}(s,t) = \int_s^t \alpha_{hj}(u)du$. Essas equações estão apresentadas em Andersen *al et.* (1993).

Consideremos ainda que

$$p_{23}(s,t) = 1 - p_{22}(s,t)$$

$$p_{13}(s,t) = 1 - p_{11}(s,t) - p_{12}(s,t)$$

A inferência pode ser baseada na maximização da verosimilhança. (Veja-se por exemplo Andersen *et al.* (1993) ou Meira Machado *et al.* (2009*)*).

3.4.1.1. Modelo de Markov em Tempo Homogéneo

Os modelos de Markov em tempo homogéneo têm sido aplicados desde há muito tempo. O modelo de Markov em tempo homogéneo é o caso mais simples de entre os modelos de Markov, sendo também o mais conhecido. Neste processo todas as intensidades de transições são assumidas como sendo constantes no tempo. Os valores reais dos riscos de transição

podem, naturalmente, ser diferentes, isto é, dependem dos estados que a transição toma. Nos modelos de Markov em tempo homogéneo é possível incluir covariáveis fixas no tempo e conhecidas, no entanto, covariáveis dependentes no tempo não podem ser consideradas.

Em modelos de Markov em tempo homogéneo, ver Meira – Machado *et al.* (2009), cada probabilidade de transição, $p_{hj}(s,t)$, depende apenas de $t-s$, ou seja,

$$p_{hj}(s,t) = p_{hj}(0, t-s).$$

Por uma questão de simplicidade de notação, só será utilizado um argumento no tempo

$$p_{hj}(t-s) = p_{hj}(0, t-s).$$

Desta forma, a equação diferencial de *Kolmogorov* tem uma solução explícita usando a decomposição da matriz de intensidades em valores e vectores próprios.

As soluções para as probabilidades de transição no modelo *illness-death* são neste caso

$$p_{11}(t) = e^{-\alpha_{12}t-\alpha_{13}t}$$

$$p_{22}(t) = e^{-\alpha_{23}t}$$

$$p_{12}(t) = \frac{\alpha_{12}(e^{-\alpha_{23}t} - e^{-\alpha_{12}t-\alpha_{13}t})}{\alpha_{12}+\alpha_{13}+\alpha_{23}}$$

Consideremos $X(\cdot)$ um processo estocástico, n indivíduos independentes, $i = 1, \dots, n$, e cada indivíduo i observado nos tempos $t_{i,0} < t_{i,1} < \cdots < t_{i,m_i}$. Observem-se os estados $x_{i,r} = X_i(t_{i,r})$, ocupados nestes momentos por cada indivíduo.

A função de verosimilhança para o modelo de Markov em tempo homogéneo é dada por

$$L = \prod_{i=1}^{n} \prod_{r=0}^{m_i-1} l_{i,r}$$

onde $l_{i,r} = p_{x_{i,r}, x_{i,r+1}}(t_{i,r+1} - t_{i,r})$ é a probabilidade de transição de cada indivíduo, i, para os estados observados (Meira-Machado *et al.* (2009)).

Embora sendo de simples e de fácil aplicabilidade, os modelos de Markov em tempo homogéneo não se adequam a todas as situações práticas, dado que considerar a intensidade de transição constante ao longo do processo, nem sempre é possível ou adequado. Assim, vários autores começam a estudar os modelos não homogéneos e os modelos semi-Markov.

Os modelos não homogéneos e de semi-Markov são comparáveis em termos de flexibilidade. A escolha entre os dois modelos depende do que se pretende modelar: a intensidade da transição desde o início do processo ou o tempo de permanência num determinado estado (Commenges (1999)).

3.5. Modelos de Regressão Multiestado

Um objectivo adicional em modelos multiestado é estudar as relações entre vários preditores e as intensidades das transições. Vários modelos têm sido utilizados na literatura para relacionar as características individuais com as taxas de intensidade, através de um possível vector de covariáveis possivelmente dependentes no tempo, Z.

Uma estratégia de simplificação comum consiste em dissociar todo o processo em vários modelos de sobrevivência, e depois ajustar separadamente intensidades de transição a todas as transições permitidas. O modelo de regressão semi-paramétrico de *hazards* proporcionais (modelo de regressão de Cox) é frequentemente utilizado sendo feitos os ajustes necessário (Meira-Machado & Roca-Pardinas (2010)) . Uma alternativa, foi proposta por Aalen originalmente para dados de sobrevivência e posteriormente usada em modelos multiestado.

Suponhamos que $\alpha_{hj,0}(\cdot)$ é a função intensidade de transição suporte entre os estados h e j, β_{hj} é o vector dos parâmetros da regressão, e Z_i é o vector das covariaveis para o indivíduo i. O modelo de regressão para modelos multiestado pode ser representado pela expressão seguinte

$$\alpha_{hji}(\cdot) = \varphi\left(\alpha_{hj,0}(\cdot), \beta_{hj}^{T} Z_i\right).$$

No caso de se considerar

$$\varphi(u(\cdot), v) = u(\cdot)e^{v}$$

obtemos o modelo de *hazard* proporcionais, descrito por

$$\alpha_{hj}(t; Z) = \alpha_{hj,0} e^{\beta^T Z}.$$

Alternativamente, escolhendo

$$\varphi(u(\cdot), v) = u(\cdot) + v,$$

temos o modelo aditivo de Aalen

$$\alpha_{hj}(t;Z) = \alpha_{hj,0} + \beta^T Z.$$

Esta equação foi proposta por Aalen e posteriormente adaptada a modelos multiestado.

3.5.1. Modelos do tipo-Cox

Os problemas de inferência em modelos multiestado podem ser dissociados em vários modelos de sobrevivência, analisando separadamente todas as intensidades em todas as transições permitidas.

Dado que no Capítulo 4 o modelo de regressão multiestado *illness –death* será aplicado a um conjunto dados reais, apresenta-se de seguida a metodologia subjacente a esse modelo.

As intensidades de transição

$$\alpha_{hj}(t;Z), 1 \leq h < j \leq 3$$

podem ser modeladas usando o modelo tipo-Cox da forma

$$\alpha_{hj}(t;Z) = \alpha_{hj,0} e^{\beta^T Z}.$$

assumindo o processo como sendo de Markov.

Para o risco de 'morte' sem doença, $\alpha_{13}(t;Z)$, o tempo de sobrevivência de pacientes que sofreram da doença são tratados como censurados no tempo de observação da doença. Os pacientes que estão vivos e livres da doença também contribuem com o tempo de sobrevivência censurados. Para a intensidade da doença, $\alpha_{12}(t;Z)$, o ponto final é o tempo de doença. Os tempos de sobrevivência dos pacientes que não se tornaram doentes são tratados como censura, quer eles estejam vivos ou tenham morrido sem terem estado doentes. Finalmente, para o modelo $\alpha_{23}(t;Z)$, que denota a intensidade da morte após a ocorrência da doença, apenas consideramos os tempos de sobrevivência (censurados ou não) truncado no tempo de doença dos indivíduos que experienciaram a doença. Note-se que os pacientes estão em risco apenas após darem entrada no estado 2. Em alguns casos, podem impor algumas restrições na função *hazard* suporte. Por exemplo, para o modelo de *illness-death*, uma

abordagem muitas vezes considerada é a de de que a finção *hazard* suporte para a transição 1 → 3 e para a transição 2 → 3 como sendo proporcionais.

Em tais casos, o modelo para essas transições é dado por

$$\alpha_{13}(t;Z) = \alpha_{13,0}e^{\beta_{13}{}^{T}Z} \qquad \text{e} \qquad \alpha_{23}(t;Z) = \alpha_{23,0}e^{\beta_{23}{}^{T}Z}.$$

O estimador para as probabilidades de transição

$$p_{hj}(s,t|Z) = P(X(t) = j|X(s) = h, Z), s \leq t \; e \; h \leq j$$

dado vector de covariáveis Z, é expresso por

$$\hat{p}_{11}(s,t|Z) = \prod_{s<u\leq t}\left(1 - \sum_{j=2}^{3} d\hat{A}_{ij}(u|Z)\right)$$

$$\hat{p}_{22}(s,t|Z) = \prod_{s<u\leq t}\left(1 - d\hat{A}_{23}(u|Z)\right)$$

$$\hat{p}_{12}(s,t|Z) = \sum_{u\leq t}\hat{p}_{11}(s,u-|Z)d\hat{A}_{12}(u|Z)\,\hat{p}_{22}(u+,t|Z)$$

onde $\hat{A}_{hj}(t|Z) = \hat{A}_{hj0}(t)e^{\hat{\beta}_{hj}{}^{T}Z}$ é o estimador da função intensidade cumulativa com $\hat{A}_{hj0}(\cdot)$ o estimador de Breslow para $\hat{A}_{hj0}(t) = \int_0^1 \alpha_{hj0}(u)du$.

Uma abordagem alternativa é usar um modelo semi-Markov em que o futuro do processo não depende do momento actual, mas sim sobre a duração do estado actual. Os modelos semi-Markov, também são denotados de modelos '*clock reset'*, porque cada vez que o paciente entra em um novo estado, o tempo é reposto a 0 (Meira-Machado *et al.* (2009)). Desta forma, os modelos de Cox semi-Markov podem ser facilmente ajustados (para o modelo de *illness-death*, a única diferença entre os modelos de Cox Markov e modelos Cox semi-Markov está na transição 2 → 3).

3.6. Introdução de Efeitos Flexíveis nas Covariáveis

Em ambos os modelos, o modelo de *hazards* proporcionais de Cox e o modelo aditivo de Aalen, o efeito de factores prognóstico é assumido com uma forma funcional linear (ou log-

linear). Contudo, se esta assumpção é violada, ela pode conduzir a conclusões estatísticas erradas: viés e uma redução do poder dos testes de significância estatística. Nos trabalhos desenvolvidos por Cox, a forma funcional incorrecta para covariáveis pode também levar a um diagnóstico de *hazards* não proporcionais. A necessidade de atenuar esta forma funcional levou a muitos desenvolvimentos em Análise de Sobrevivência. Usando os modelos tipo-Cox descritos anteriormente, a implementação destes modelos no âmbito dos modelos de multiestado pode ser facilmente considerada. No entanto, a implementação dos efeitos de uma covariável não linear noutras abordagens de multiestado não é simples.

No contexto do modelo de regressão de Cox (*hazards* proporcionais) várias abordagens foram propostas para o problema de testar a linearidade. Uma abordagem geral para o modelo de *hazards* proporcionais com um efeito arbitrário de covariáveis é dada por

$$\alpha_{hj}(t;Z) = \alpha_0(t)e^{f(Z)},$$

onde $f(Z)$ é assumida como sendo uma função suavizadora de Z. Tibshirani & Hastie (1987) consideram uma função não paramétrica para este problema. Apesar de este modelo não garantir o risco como log-linear em Z, é difícil interpretar a influência de uma única covariável no tempo de sobrevivência. Ainda, quando Z tem demasiadas componentes esta abordagem está sujeita a problemas de dimensionalidade (para mais detalhes consultar Meira-Machado *et al*. (2009)). Este problema pode ser contornado reduzindo a dimensionalidade através de um modelo de regressão aditivo, em que se expressa a *log hazard* como uma função aditiva de cada covariável

$$\alpha_{hj}(t;Z) = \alpha_0(t)e^{\sum_{j=1}^{p} f_j(Z_j)},$$

onde $f_j(\cdot), j = 1, \ldots, p$ são funções de covariáveis suaves não especificadas. Estes modelos têm vindo a ser estudados por vários autores usando técnicas não paramétricas. *Splines* baseados em métodos de suavização foram considerados por exemplo por Tibshirani e Hastie (1987).

Este modelos podem ser usados para testar a presença dos efeitos não lineares e para identificar a forma funcional correcta.

Capítulo 4

Aplicação a Dados de Transplante do Coração

4.1. Software

No decorrer do estudo serão utilizados os programas estatísticos R e *BayesX* .

R-Project é uma implementação *Open Source* do software estatístico S-Plus. Utiliza uma linguagem muito semelhante ao C e pode ser pode ser obtido, gratuitamente, a partir do endereço http://www.r-project.org.

Relativamente ao software estatístico R utilizam-se diversas funções da *package tdc.msm e p3state.msm* desenvolvidas por vários autores para ambiente R. O *download* da *package tdc.msm* e a base de dados *stanford* presente neste estudo pode ser realizado gratuitamente a partir de http://www.mct.uminho.pt/lmachado/Rlibrary. Desenvolvida por Meira-Machado, Cadarso-Suárez e Uña-Álvarez a package *tdc.msm* permite a análise de dados de sobrevivência utilizando modelos multiestado. Para modelos multiestados são considerados modelos diferentes, permitindo escolher entre modelos de Markov e semi-Markov, bem como a utilização de modelos homogéneos e não homogéneos. *Outputs* gráficos incluindo estimativas de funções de sobrevivência, estimativas de probabilidades de transição e curvas suavizadas *log hazard* para covariáveis contínuas podem também ser obtidos.

A *package p3state.msm* desenvolvida por Meira-Machado e Roca-Pardiñas permite analisar dados de sobrevivência em modelos de *illness-death* e modelos de três estados. O *package p3state.msm* permite ainda estimar modelos de regressão semi-paramétricas e implementar estimadores não paramétricos para probabilidades de transição. No site http://cran.r-

project.org/web/packages/p3state.msm/ pode-se proceder ao download desta *package* gratuitamente. A base de dados, *heart2*, associada a este *software* encontra-se no mesmo *site*.

Desenvolvida por Arthur Allignol a *package etm* fornece estimtivas e gráficos para as probabilidades de transição e os seus gráficos em qualquer modelo multiestado em tempo homogéneo, com um número finito de estados. Pode fazer-se o *download* desta *package* em http://cran.r-project.org/web/packages/etm/index.html.

O software *BayesX* foi desenvolvido para estimar modelos aditivos generalizados mistos englobando diversos tipos de modelos complexos de regressão (Brezger *et al.* (2005)). *BayesX* tem distribuição gratuita e pode ser encontrado na plataforma *MicroSoft Windows* (www.statistik.lmu.de/~bayesx/).

Ao longo deste trabalho serão colocados os comandos utilixados para a obtenção dos vários resultados e representações gráficas.

4.2. Base de Dados Transplante do Coração de Stanford

A base de dados *Stanfort Heart Transplant* é uma base de dados clássica de análise de sobrevivência com covariáveis. Os dados reportam-se ao programa de transplante de coração de Stanford no período que decorrido entre Outubro de 1967 e Abril de 1974.

Na Tabela 4.1 ilustra-se o formato dos dados originais de Crowley e Hu (1977). Há 103 pacientes, apresentando-se apenas os dados para os 10 primeiros indivíduos.

Pacient	Date of birth	Date of acceptance	Date of transplant	Date last seen	Deado=1 Alive=0	Previous surgery
1	1/10/37	11/15/68		1/3/68	1	0
2	3/2/16	1/2/68		1/7/68	1	0
3	9/19/13	1/6/68	1/6/68	1/21/68	1	0
4	12/23/27	3/28/68	5/2/68	5/5/68	1	0
5	7/28/47	5/10/68		6/15/68	1	0
6	11/8/13	6/13/68		5/27/68	1	0
7	8/29/17	7/12/68	3/31/68	5/17/68	1	0
8	3/27/23	8/1/68		9/9/68	1	0
9	6/11/21	8/9/68		11/1/68	1	0
10	2/9/26	8/11/68	8/22/68	10/7/68	1	0

Tabela 4.1- Formato original dos dados transplante de coração de Stanford (primeiros 10 pacientes)

Existem dois tipos de pacientes: aqueles que receberam um transplante, e os que foram incluídos no estudo e ficaram a aguardar dador compatível para receberem transplante mas tal não viria a acontecer.

Um aspecto crucial da análise é a transformação dos dados originais num formato consistente com a formulação do processo de contagem. Esta base de dados, denotada por ***stanford***, pode ser encontrada na *package* survival assim como na *package tdc.msm*.

Na Tabela 4.2, ilustra-se os dados transformados para os primeiros 10 pacientes. Note-se que as 10 linhas originais tornaram-se 14; cada um dos pacientes que recebeu um transplante, ou seja, indivíduos {3, 4, 7, 10} têm agora duas linhas em vez de uma. No presente exemplo, para pacientes transplantados, temos dois registos desse tipo para cada paciente, que corresponde ao intervalo de tempo entre a entrada no estudo e a realização do transplante, e o outro para o intervalo entre o transplante e a morte ou fim do estudo. Para os pacientes que não receberam um transplante existe apenas uma única linha.

	Id	***start***	***stop***	***event***	***transplant***	***age***	***year***	***surgery***
1	1	0.0	50.0	1	0	-1.715.537.303	0.12320329	0
2	2	0.0	6.0	1	0	383.572.895	0.25462012	0
3	3	0.0	1.0	0	0	629.705.681	0.26557153	0
4	3	1.0	16.0	1	1	629.705.681	0.26557153	0
5	4	0.0	36.0	0	0	-773.716.632	0.49007529	0
6	4	36.0	39.0	1	1	-773.716.632	0.49007529	0
7	5	0.0	18.0	1	0	-2.721.423.682	0.60780287	0
8	6	0.0	3.0	1	0	659.548.255	0.70088980	0
9	7	0.0	51.0	0	0	286.926.762	0.78028747	0
10	7	51.0	675.0	1	1	286.926.762	0.78028747	0
11	8	0.0	40.0	1	0	-265.023.956	0.83504449	0
12	9	0.0	85.0	1	0	-0.83778234	0.85694730	0
13	10	0.0	12.0	0	0	-549.760.438	0.86242300	0
14	10	12.0	58.0	1	1	-549.760.438	0.86242300	0

Tabela 4.2 - Formato transformado dos dados transplante de coração de Stanford (***stanford***) (primeiros 10 pacientes)

No conjunto de dados há 69 transplantes no total de 103 pacientes pelo que nos dados transformados se tem 172 linhas $(2 \times 69 + 44 = 172)$. No fim do referido período 28 pacientes contribuíram com tempos de sobrevivência censurados e 75 pacientes haviam falecido (45 com transplante e 30 sem terem recebido transplante).

Na Tabela 4.2 estão disponíveis várias informações para cada indivíduo. A variável *age*, expressa em anos, é uma variável transformada tal que *age* = (idade do indivíduo quando incluído no estudo) – 48, desta forma 0 significa que o indivíduo tem 48 anos quando é incluído no estudo. A variável *year* significa o ano de entrada no estudo, expressa por (data de entrada em estudo em dias desde 10/01/1967)/365.25. Esta variável é medida em anos

(*year*=0 significa que o indivíduo é aceite no estudo a 10/01/1967). Tal como referido anteriormente, para um indivíduo que foi sujeito a um transplante existem 2 linhas: a primeira linha proporciona informação do tempo até ao transplante, e a segunda para o tempo que decorre desde o transplante até à morte ou fim de estudo. Aqui *transplant* é uma covariável dependente no tempo (0 - Não; 1 - Sim). A variável *surgery* indica se o indivíduo teve ou não alguma cirurgia prévia para colocação de *bypass* (0 - Não; 1 - Sim), as variáveis *start* e *stop* que apresentam informação sobre o intervalo de tempo que definem os eventos (transplante e morte) de cada indivíduo em estudo, a variável *event* indica se o indivíduo registou o evento de interesse, neste caso a morte, ou apresentou censura (0 - Censura; 1 - Evento de interesse), finalmente a variável *id* que fornece a identificação do paciente. Por exemplo, as observações nas linhas 3 e 4 representam o mesmo indivíduo. Este indivíduo foi sujeito a um transplante depois apenas 1 dia depois de entrar em estudo e faleceu 15 dias depois.

Assim, as variáveis que compõem a base de dados, e a partir das quais se acede à referida informação, são: *start, stop, event, year, surgery, transplant, id*. De realçar, que a covariável *transplant* é a única covariável dependente no tempo, enquanto as restantes covariáveis são fixas.

Na análise aos dados será também utilizada a base de dados ***heart2*** presente na *package p3state.msm*, referente ao transplante de coração de Stanford, mas apresentando um formato diferente dos anteriores. Na Tabela 4.3 ilustra-se os dados para os primeiros 10 pacientes na base de dados *heart2*.

	times1	***delta***	***times2***	***time***	***status***	***age***	***year***	***surgery***
1	50.0	0	0.0	50	1	-1.715.537.303	0.12320328	0
2	6.0	0	0.0	6	1	383.572.895	0.25462012	0
3	1.0	1	15.0	16	1	629.705.681	0.26557153	0
4	36.0	1	3.0	39	1	-773.716.632	0.49007529	0
5	18.0	0	0.0	18	1	-2.721.423.682	0.60780287	0
6	3.0	0	0.0	3	1	659.548.255	0.70088980	0
7	51.0	1	624.0	675	1	286.926.762	0.78028747	0
8	40.0	0	0.0	40	1	-265.023.956	0.83504449	0
9	85.0	0	0.0	85	1	-0.83778234	0.85694730	0
10	12.0	1	46.0	58	1	-549.760.438	0.86242300	0

Tabela 4.3- Formato transformado dos dados Transplante Coração de Stanford ***(heart2)*** (primeiros 10 pacientes)

Esta base de dados apresenta 8 variáveis. A variável *times1* representa o tempo de transplante/ censura, a variável *delta* é o indicador se o paciente recebeu transplante (*delta*=1) ou nunca chegou a receber (*delta*=0). A variável *times2* apresenta o tempo até ao evento de interesse,

morte, contado a partir do momento de transplante. A variável que indica o tempo total que o paciente esteve em estudo é a variável *time*, que é dado pela soma das variáveis *time1* e *time2*. O indicador de censura é a variável *status* (0 - Censura; 1 - Evento de interesse). Finalmente, as variáveis *age*, *year* e *surgery* são as mesmas descritas anteriormente.

Para estimar as probabilidades de transição para o modelo de Markov em tempo homogéneo será necessário recorrer a uma nova transformação dos dados. Na Tabela 4.4 apresenta-se o novo formato para o conjunto de dados.

	id	***entry***	***exit***	***from***	***to***	***age***	***year***	***surgery***
1	1	0.0	50.0	1	3	-1.715.537.303	0.12320328	0
2	2	0.0	6.0	1	3	383.572.895	0.25462012	0
3	3	0.0	1.0	1	2	629.705.681	0.26557153	0
4	3	1.0	16.0	2	3	629.705.681	0.26557153	0
5	4	0.0	36.0	1	2	-773.716.632	0.49007529	0
6	4	36.0	39.0	2	3	-773.716.632	0.49007529	0
7	5	0.0	18.0	1	3	-2.721.423.682	0.60780287	0
8	6	0.0	3.0	1	3	659.548.255	0.70088980	0
9	7	0.0	51.0	1	2	286.926.762	0.78028747	0
10	7	51.0	675.0	2	3	286.926.762	0.78028747	0

Tabela 4.4- Formato transformado dos dados Transplante Coração de Stanford ***(heart3)*** (primeiros 10 pacientes).

Nesta estrutura de dados, as variáveis *from* e *to* indicam, respectivamente, o estado em que o paciente se encontra e o estado para onde este se deslocou. A transição (*from*=1 *to*=3) denota a transição da mortalidade sem o transplante, a transição (*from*=1 *to*=2) indica a transição de transplante e a transição (*from*=2 *to*=3) a mortalidade de transição após o transplante. Desta forma, como podemos observar na Tabela 4.4, o individuo 3 foi sujeito a transplante ocorendo as transições *from*=1 *to*=2 e *from*=2 *to*=3. As variáveis *id*, *entry*, *exit*, *age*, *year* e *surgery* denotam, respectivamente, *id, start, stop, age*, *year* e *surgery* descritas anteriormente no formato *stanford* da base de dados.

4.2. Análise Exploratória dos Dados

Como em qualquer análise estatística, a análise descritiva dos dados deve ser o primeiro passo no estudo de tempos de vida. Para se proceder à caracterização da base de dados inicia-se o

presente estudo com uma breve análise descritiva das variáveis. É particularmente pertinente a análise das covariáveis anteriormente referidas.

Na Tabela 4.5 apresentam-se os valores de algumas medidas de localização e medidas de dispersão para as covariáveis quantitativas *age* e *year*.

Variável	Média	Mediana	Amplitude	Desvio Padrão
age	-2.4840	-0.1136	-39.2142 ; 16.4079	9.4199994
year	3.45329	3.75086	0.04928 ; 6.47228	1.8249273

Tabela 4.5 - Medidas de localização e de dispersão para os dados.

Note que à variável idade é necessário acrescentar 48 anos, dado que na base de dados, para esta variável, 0 significa que o indivíduo tinha 48 anos quando foi incluído no estudo. Desta forma, a média das idades de entrada no estudo é, aproximadamente, 45 anos. As idades dos indivíduos, quando estes se submetem à experiência, varia dos 8 aos 64 anos.

Nesta parte do estudo recorre-se à base de dados *heart2* presente no software estatístico *p3state.msm*.

O gráfico da Figura 4.1 faculta uma análise mais detalhada da variável age. É interessante realçar que os 50% valores centrais se situam entre, aproximadamente, -6 e 4 (ou seja, entre 42 e 52 anos), sugerindo uma maior incidência de transplantes em pessoas de média idade.

Importa também salientar a existência de outliers, valores esses considerados atípicos por serem assaz menores do que os restantes. No R, executa-se o comando

```
R>library(p3stat.msm
R>data(heart2)
R>attach(heart2)
R>boxplot(age+48, col=c("lightgrey"))
```

A comparação entre o ano de aceitação de um paciente no estudo e o seu estatuto pode ser feita recorrendo a caixas-com-bigodes em paralelo, conforme se observa na Figura 4.2.

Atentando agora no gráfico da Figura 4.2 verifica-se que as medianas são bastante diferentes, evidenciando menores valores da variável *year* para o nível 'morte' da variável *status*. Esta constatação sugere melhorias (traduzidas pelo menor número de mortos) com o decorrer dos anos. Tal pode dever-se ao desenvolvimento de novos e melhor métodos de tratamento de

doenças cardíacas, durante o decorrer do período de observação. Outra explicação pode ser dada por uma possível alteração dos critérios de admissão dos pacientes no estudo.

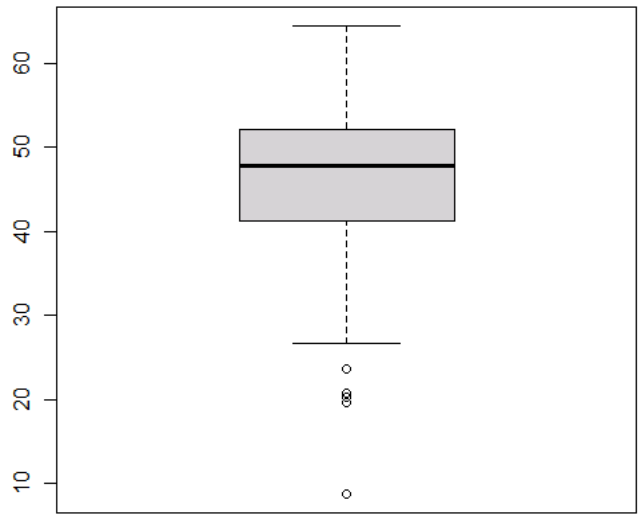


Figura 4.1- Caixa-com-bigodes para a variável *age+48*.

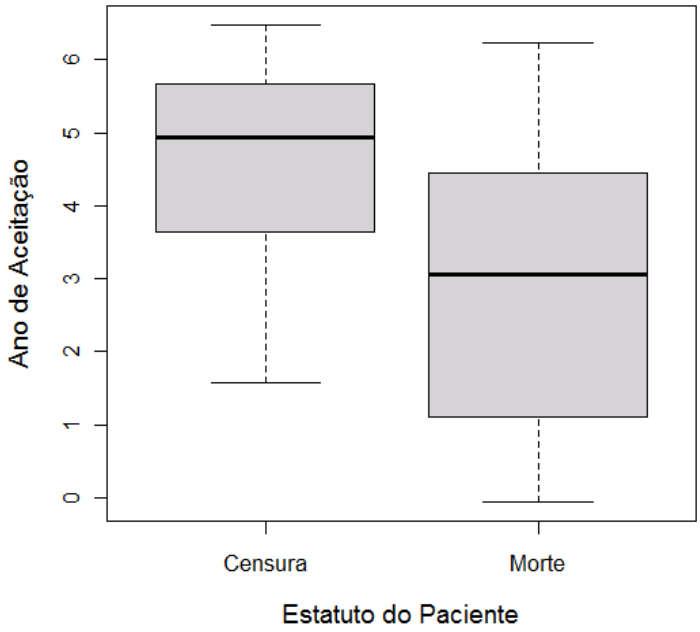


Figura 4.2 - Caixa-com-bigodes para a variável Ano de Aceitação em função de Estatuto.

O gráfico da Figura 4.2 pode ser obtido e executando-se no R o seguinte comando.

```
R>boxplot(year~status, cex.lab=1.2, cex.axis=1.1, col=c("lightgrey","lightgrey"),
names=c("Censura","Morte"), ylab="Ano de Aceitação", xlab="Estatuto do Paciente")
```

Na Tabela 4.6 apresentam-se os valores de algumas medidas de localização e medidas de dispersão para o tempo global de sobrevivência. Em ambiente R,

```
R>sd(time)
```

	Mínimo	***1º Quartil***	***Mediana***	***Média***	***3º Quartil***	***Máximo***	***Desvio Padrão***
Tempo	1	33.5	90	310.2	412.5	1800	428.3

Tabela 4.6- Algumas medidas de localização e medidas de dispersão para o tempo global de sobrevivência

Por último compare-se o tempo de sobrevivência para os pacientes que não realizaram transplante e para os que realizaram, sendo que para estes últimos será considerado apenas o tempo após a realização do mesmo.

No R executa-se o comando

```
R>boxplot(times1, times2, cex.lab=1.2, cex.axis=1.1,col=c("lightgrey","lightgrey"),
names=c("Sem transplante","Com transplante"))
```

No gráfico da Figura 4.3 observa-se uma grande diferença no tempo de vida entre os pacientes que realizaram transplante e os que não realizaram. Em particular, evidencia-se um maior tempo de sobrevivência para os pacientes transplantados. Convém realçar que a comparação destes tempos deve ser analisada com cuidado pois algumas destas observações são censuradas pela direita.

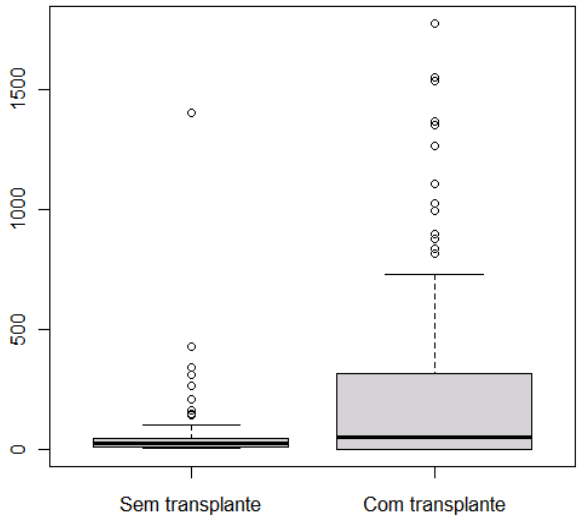


Figura 4.3 - Caixa-com-bigodes para os tempos de sobrevivência para indivíduos que realizaram.

Em relação a tempos de sobrevivência, uma análise mais detalhada envolvendo covariáveis, será desenvolvida nas secções seguintes.

4.3. Modelos de Análise de Sobrevivência

4.3.1. Função de Sobrevivência

Dada a existência de censura, os dados de sobrevivência são resumidos de forma conveniente através de estimativas da função de sobrevivência e da função de risco (ou função *hazard*). No estudo que se segue, apresentaremos o estimador de *Kaplan-Meier* para a função de sobrevivência aplicado à base de dados de transplante do coração de *Stanford*. Trata-se de métodos não paramétricos, porque a estimação é feita sem que se faça nenhuma suposição sobre a distribuição de probabilidade do tempo de sobrevivência. Iremos, também, utilizar alguns métodos não paramétricos para a comparação de curvas de sobrevivência, nomeadamente o teste *log-rank.*

Em ambiente R tal teste pode ser efectuado recorrendo à função *survfit* que oferece várias opções e funcionalidades.

Comece-se por estimar e representar a curva de sobrevivência. Para mais informações relativas à curva de sobrevivência *Kaplan-Meier* global acima estimada pode consultar-se no Anexo I.

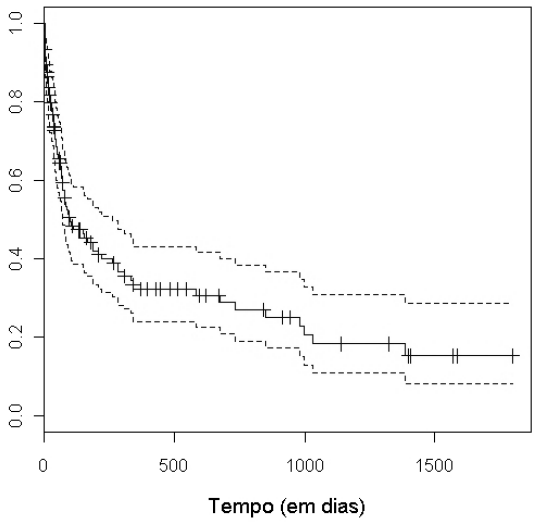


Figura 4.4 - Curva de sobrevivência.

No gráfico da Figura 4.4 representa-se a estimativa *Kaplan- Meier* da curva de sobrevivência para os dados. Conforme se observa para valores baixos para a variável tempo verifica-se a ocorrência de muitos eventos (assinalados pelos pequenos traços verticais na curva). Tal implica, como também se observa, um decrescimento acentuado da curva de sobrevivência nos menores valores de tempo. Como se pode verificar as curvas de sobrevivência apresentam na realidade uma forma de escada, correspondendo cada degrau a um novo evento. Quando o número de indivíduos em risco é maior, estes degraus são mais pequenos e as curvas ficam com um traço mas regular, como podemos observar na Figura 4.4. Repare-se que para valores superiores de tempo a curva parece estabilizar.

Examinada a curva de sobrevivência para os dados analisam-se seguidamente as curvas de sobrevivência para grupos formados pelas covariáveis qualitativas, ou seja, para as variáveis transplante (*delta*) e *surgery.*

Um olhar atento sobre os gráficos da Figura 4.5 sugere a existência de diferenças significativas para as curvas de sobrevivência considerando o grupo dos pacientes que foram transplantados e o grupo dos que não foram. De facto, a presença de paralelismo para estes dois grupos evidencia existir diferenças significativas nas respectivas curvas de sobrevivência. No entanto, relativamente às curvas de sobrevivência quanto aos pacientes com/sem cirurgia prévia parecem evidenciar-se algumas diferenças, sendo que a realização de cirurgia prévia para introdução de *bypass* parece estar associada a um aumento do tempo de sobrevivência.

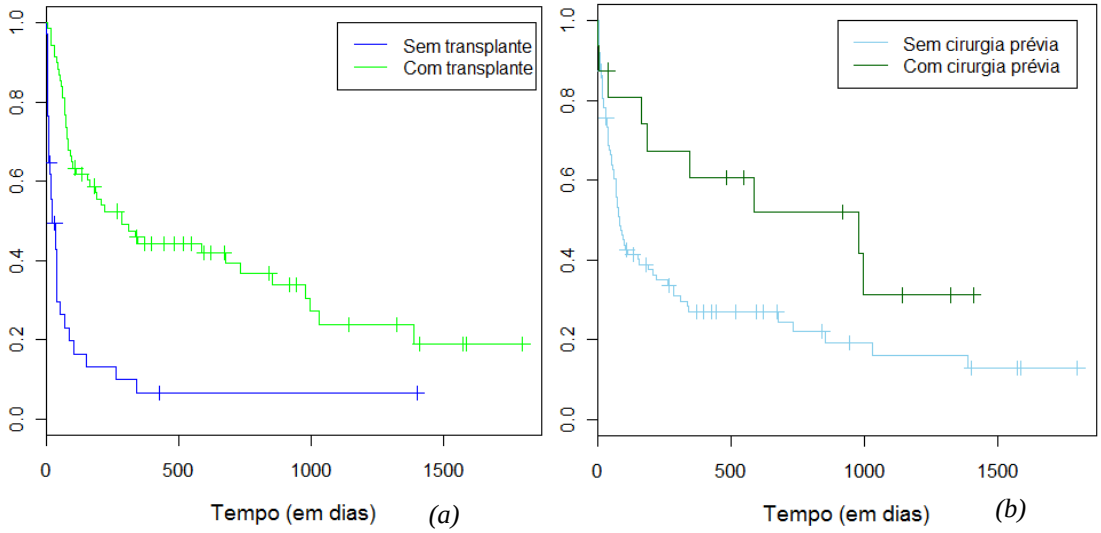


Figura 4.5-Curvas de sobrevivência *Kaplan-Meier* em função de: *(a) Transplante (delta); (b) Surgery*.

Os gráficos da Figura 4.5 podem ser obtidos no R executando os seguintes comandos.

```
R>kmt=survfit(Surv(time,status)~delta,data=heart2)
R>plot(kmt, col=c("blue","green"), cex=1.2, cex.axis=1.1, xlab="Tempo (em dias)", cex.lab=1.2)
legend(x=c(1100,1850), y=c(0.83,1), legend=c("Sem transplante","Com transplante"), lty=1, col=c("blue","green"))
R>kms=survfit(Surv(time,status)~surgery,data=heart2)
plot(kms, col=c("skyblue","darkgreen"), cex=1.2, cex.axis=1.1, xlab="Tempo (em dias)", cex.lab=1.2)
legend(x=c(900,1800), y=c(0.83,1), legend=c("Sem cirurgia prévia", "Com cirurgia prévia"), lty=1, col=c("skyblue","darkgreen"))
```

Depois das funções de sobrevivência estimadas e representadas graficamente será importante testar se existem diferenças significativas na sobrevivência desses diversos grupos.

Existem vários métodos não paramétricos, para detectar diferenças entre as distribuições dos tempos de vida de vários grupos, visto a estatística de teste depender unicamente das ordens das observações. Um desses métodos, e que será usado neste estudo, consiste em efectuar o teste (de hipóteses) de *log-rank*, segundo o qual se tem como hipótese nula a inexistência de diferenças das curvas de sobrevivência contra a hipótese alternativa de existência de diferenças nas curvas. Em ambiente R tal teste pode ser efectuado recorrendo à função *survdiff*. Realize-se agora o teste de hipótese *de log-rank* para detectar possíveis diferenças

significativas nas curvas de sobrevivência para os dois grupos da variável *surgery* e transplante (*delta*).

Na Tabela 4.7 apresentam-se os valores-prova para os dois testes realizados.

Variável	Estatística de teste	Grau de Liberdade	Valor prova
surgery	4.4	1	0.035
transplante (*delta*)	33.2	1	< 0.001

Tabela 4.7 - Teste de *log-rank*.

Relativamente à variável *surgery* constatam-se diferenças na curva de sobrevivência para os dois grupos (com ou sem cirurgia prévia) ao nível de significância de 5%, sendo esta observação consistente com as ilações gráficas retiradas anteriormente. Para a variável transplante (*delta*), como pode observar-se o valor de prova é inferior a 0.001 revelando forte evidência estatística de que as curvas de sobrevivência diferem significativamente para os dois grupos (os que foram transplantados e os que não foram). Ora, este resultado é consistente com o verificado na análise do gráfico (a) da Figura 4.5. Os referidos testes podem obter-se no R conforme se descreve abaixo.

```
R>survs=survdiff(Surv(time, status)~surgery, rho=0)
R>survt=survdiff(Surv(time, status)~delta, rho=0)
```

Expostas as principais conclusões acerca da estimação da função de sobrevivência de Kaplan-Meier e respectiva aplicação aos dados importa salientar alguns pontos fulcrais.

Efectivamente, como disposto anteriormente, o estimador Kaplan-Meier providencia várias vantagens na realização de um estudo do campo de análise de sobrevivência. Não obstante, este estimador apresenta algumas limitações. Nomeadamente, é maioritariamente descritivo, não controla covariáveis (quantitativas), os preditores têm de ser categóricos e não incorpora variáveis dependentes no tempo.

4.3.2. Regressão de Cox

Referidos no Capítulo 2 os aspectos essenciais sobre o modelo de regressão de *Cox* proceda-se à sua aplicação aos dados, utilizando o *software* R, iniciando o estudo com modelos de

regressão univariados, e seguido-se os modelos de regressão multivariados. Para tal invocar-se-á a função *coxph*. Assim sendo, ajustem-se vários modelos de regressão de *Cox* para os dados h*eart2*.

Na Tabela 4.8 apresentam-se informações várias acerca dos modelos ajustados. A título exemplificativo apresenta-se abaixo o código necessário, em R, para dois desses modelos, e para o cálculo do AIC.

```
R> heart1=coxph(Surv(time,status)~age,data=heart2)
 R>AIC1= -2*heart1$loglik[2]+2
 R>heart2=coxph(Surv((time,status)~year,data=heart2)
 R>AIC5=-2*heart2$loglik[2]+4
```

A escolha do melhor modelo de regressão de Cox é uma difícil tarefa pois raramente existe um único melhor modelo.

Por conseguinte, é necessário estabelecer uma estratégia que seja um compromisso entre a eleição de um modelo que proporcione um bom ajuste aos dados e, por outro lado, seja fácil de interpretar os dados. Para tal é habitual utilizar os procedimentos *stepwise*. Porém, uma vez que a base de dados em estudo consta apenas de quatro covariáveis optou-se, também, por analisar o AIC de todos os modelos de regressão de *Cox* possíveis (fazendo todas as combinações possíveis). Os resultados figuram na Tabela 4.8.

Em todos os modelos apresentados, a influência da variável age no risco é positiva (*Hazard Ratio*>1º efeito de risco), enquanto os efeitos das variáveis year e surgery são ambos negativos (*Hazard Ratio* <1; efeito protector) (ver Tabela 4.8).

O critério de informação de *Akaike* é um critério de selecção de um modelo. Um valor baixo para AIC é considerado como representativo de um melhor ajustamento e, assim, na selecção de modelos deve-se ter como objectivo a minimização de AIC. Ao analisar os modelos que apresentam menor AIC, vemos que o efeito do transplante (variável *delta*) conduz a uma pequena redução no risco, mas sem alcançar significância estatística. Conforme se observa na Tabela 4.8 o modelo com menor AIC é aquele que tem *age, year* e *surgery* como covariáveis.

Contudo, e atentando nos sucessivos valores-prova obtidos para a covariável suspeita-se que a variável *surgery* não é importante no modelo.

	Estimativa	*age*	*year*	*surgery*	**transplante (*delta*)**	*AIC*
heart1	$\hat{\beta}(SE)$ p-value	0.0307(0.0143) 0.031				593.0735
heart2	$\hat{\beta}(SE)$ p-value		-0.191 (0.07) 0.0064			590.7184
heart3	$\hat{\beta}(SE)$ p-value			-0,74 (0.359) 0.039		593.1680
heart4	$\hat{\beta}(SE)$ p-value				0.127 (0.301) 0.67	598.0629
heart5	$\hat{\beta}(SE)$ p-value	0.0265 (0.0137) 0.054	-0.1784 (0.0704) 0.011			588.5850
heart6	$\hat{\beta}(SE)$ p-value	0.0307 (0.0136) 0.024		-0.7728 (0.462) 0.032		589.5268
heart7	$\hat{\beta}(SE)$ p-value	0.03074 (0.0145) 0.034			-0.00418 (0.312) 0.99	595.0733
heart8	$\hat{\beta}(SE)$ p-value		-0.162 (0.070) 0.021	-0.598 (0.366) 0.100		589.6840
heart9	$\hat{\beta}(SE)$ p-value		-0.191 (0.0701) 0.0064		0.123 (0.3031) 0.6800	592.5518
heart10	$\hat{\beta}(SE)$ p-value			-0.749 (0.360) 0.037	0.158 (0.297) 0.590	594.8803
heart11	$\hat{\beta}(SE)$ p-value	0.0271 (0.0134) 0.043	-0.1463 (0.0704) 0.038	-0.6376 (0.3670) 0.082		587.1323
heart12	$\hat{\beta}(SE)$ p-value	0.0268 (0.0141) 0.057	-0.1787 (0.0704) 0.011		-0.0308 (0.3175) 0.920	590.5756
heart13	$\hat{\beta}(SE)$ p-value	0.0305 (0.0139) 0.028		-0.7733 (0.3597) 0.032	0.0161 (0.3086) 0.960	591.5240
heart14	$\hat{\beta}(SE)$ p-value		-0.162 (0.070) 0.021	-0.607 (0.367) 0.098	0.148 (0.299) 0.620	591.4386
heart15	$\hat{\beta}(SE)$ p-value	0.0272 (0.0137) 0.048	-0.1463 (0.0705) 0.038	-0.6372 (0.3672) 0.083	-0.0103 (0.3138) 0.974	589.1312

Tabela 4.8-Resumo de variadas combinações de modelos de regressão de Cox.

No que diz respeito aos métodos *stepwise* procedeu-se no R à utilização da função *stepAIC* disponível na biblioteca *MASS*. Empregou-se o método *backward elimination* e o método composto, sendo que, neste último caso, partiu-se de um modelo que parecia ser próximo de um modelo adequado (com *age* e *year* como covariáveis). Os resultados alcançados foram consistentes com os anteriores. Assim um modelo que descreve adequadamente os dados tem como covariáveis *age*, *year* e *surgery*.

Em ambiente R basta executar os comandos seguintes.

```
R>modelo=coxph(Surv(time,status)~ age+year, data=heart2)
R>resultado<-stepAIC(modelo, scope=list(upper =~age+year+surgery+delta, lower = ~1))
R>modelo1=coxph(Surv( time,status )~age+year+surgery+delta, data=heart2)
R>resultado=stepAIC(modelo1)
```

O output de cada um dos comandos acima contendo informação relevante para a compreensão do método de selecção empregue apresentam-se de seguida.

Backward elimination			**Método Composto**		
Start: AIC=588.59			Start: AIC=555.46		
Surv(time, status) ~ age + year			Surv(time, status) ~ age + year + surgery + delta		
	Df	AIC		Df	AIC
+ delta	1	554.21	- surgery	1	554.21
+ surgery	1	587.13	<none>		555.46
			- year	1	556.39
<none>		588.59			
- age	1	590.72	- age	1	571.73
- year	1	593.07	- delta	1	587.13
			Step: AIC=554.21		
Step: AIC=554.21					
			Surv(time, status) ~ age + year + delta		
Surv(time, status) ~ age + year + delta				Df	AIC
	Df	AIC	<none>		554.21
<none>		554.21	- year	1	555.78
+ surgery	1	555.46	- age	1	570.66
- year	1	555.78	- delta	1	588.59
- age	1	570.66			
- delta	1	588.59			

Tabela 4.9- *Outputs* dos métodos *stepwise*

Na Tabela 4.10 apresentam-se informações relevantes referentes ao modelo considerado adequado.

	β	e^{β}	Lim. Inf. I.C. 95%	Lim. Sup. I.C. 95%	Valor - Prova
age	0.0271	1.027	1.001	1.055	0.043
year	-0.1463	0.864	0.753	0.992	0.038
surgery	-0.06376	0.529	0.257	1.085	0.082
Likelihood ratio test = 15.1			df=3	p=0.00172	

Tabela 4.10- Estimativas de coeficientes de regressão de *Cox*

A observação da Tabela 4.10 permite concluir:

• A variável *age* é significativa ao nível de significância de 5%. Atentando no valor de e^{β} conclui-se que o risco de um indivíduo quando comparado com um indivíduo um ano mais novo é aproximadamente 2,7% maior (sendo os valores das restantes covariáveis iguais!);

• A covariável *year* proporciona um coeficiente estimado negativo, significando que o risco de morte é menor à medida que esta variável aumenta, ou seja, o risco de morte diminui à medida que a data de entrada no estudo aumenta. De facto, o risco de morte é 86.4% do de um indivíduo com uma unidade a menos no ano de aceitação. Repare-se que esta é significativa ao nível de significância de 5%;

• A covariável s*urgery* apenas é significativa no modelo ao nível de significância de 10%. Note-se como o valor 1 pertence ao intervalo de confiança a 95% para e^{β}. No entanto, apesar desse facto optou-se pela sua manutenção no modelo devido à melhoria introduzida ao nível de AIC;

• Como a categoria de referência para a variável s*urgery* é 0 o valor 0.529 significa que o risco de morte de um indivíduo que realizou cirurgia prévia quando comparado com um indivíduo que realizou é aproximadamente metade do de um paciente que não realiza cirurgia prévia.

4.3.3. Efeito Linear versus Efeito Não Linear

A abordagem de modelo de *hazards* proporcionais providencia apenas estimativas constantes para o efeito de uma covariável contínua para todo o período de estudo, o que pode ser considerada uma desvantagem do modelo. Interessa por isso estudar se o efeito de determinada covariável contínua é ou não linear.
Recorrendo ao *software* R, pode-se fazer uso da função *pspline* para averiguar o efeito das variáveis *age* e *year* é linear. Primeiramente será usado o mecanismo automático de selecção do número de graus de liberdade (Critério de Informação de *Akaike*), para $df = 0$ na função *pspline*.

```
R>fit1=coxph(Surv(time,status)~pspline(age,0)+year+factor(surgery,levels=c(1,0)),data=heart
2)
R>fit2=coxph(Surv( time,status )~age+pspline(year,0)+factor(surgery,levels=c(1,0)),
data=heart2)
```

Variável	*Efeito*	*Estatística de Teste*	*Graus de Liberdade*	*Valor-Prova*
Age	Linear	3.12	1.00	0.077
	Não Linear	6.85	4.94	0.230
Year		3.44	1.00	0.064
surgery		2.50	1.00	0.110

Tabela 4.11- Averiguação de efeito linear ou não linear da variável *age*.

Variável	*Efeito*	*Estatística de Teste*	*Graus de Liberdade*	*Valor-Prova*
Age		5.24	1.00	0.022
Year	Linear	5.90	1.00	0.015
	Não Linear	9.37	2.93	0.023
surgery		4.23	1.00	0.40

Tabela 4.12 - Averiguação de efeito linear ou não linear da variável *year*.

Na Tabela 4.11 verifica-se que o efeito da covariável *age* deve ser considerado linear. De facto, o valor-prova 0.230 permite concluir que não existet vantagem em introduzir a variável *age* como efeito não linear.

A observação da Tabela 4.12 permite constatar que o efeito da variável *year* deve ser considerado como não linear. De facto, como se tem valor-prova igual a 0.023 significa que esta variável é significativa no modelo quando introduzida com efeito não linear.

Tendo em atenção as conclusões anteriores, particularmente as referentes à covariável *year*, apresenta-se na Figura 4.6 o gráfico para o efeito não linear dessa variável. Repare-se que a curva representada nessa mesma figura consiste na função de risco suavizada em função da variável *year*. Pode observar-se uma diminuição significativa do risco até ao primeiro ano e meio, aproximadamente, verificando-se posteriormente uma estabilização do risco com pequenas diminuições.

A obtenção do gráfico da Figura 4.6 passa pela execução dos comandos abaixo, no software R.

```
R>fit=coxph(Surv(time,status)~pspline(year,0), data=heart2)
R>tem=predict(fit, type="terms", se.fit=T)
R>tmat=cbind(tem$fit, tem$fit-1.96*tem$se.fit, tem$fit+1.96*tem$se.fit)
R>jj=match(sort(unique(heart2$year)),heart2$year)
R>matplot(heart2$year[jj],tmat[jj,],type="l",xaxt="n", xlab="year", lty=c(1,2,2), ylab="log
R>hazard ratio", col=c("black"), lwd=2, cex.lab=1.2)
R>xx=c(0,1,2,3,4,5,6,7)
R>axis(1,xx)
```

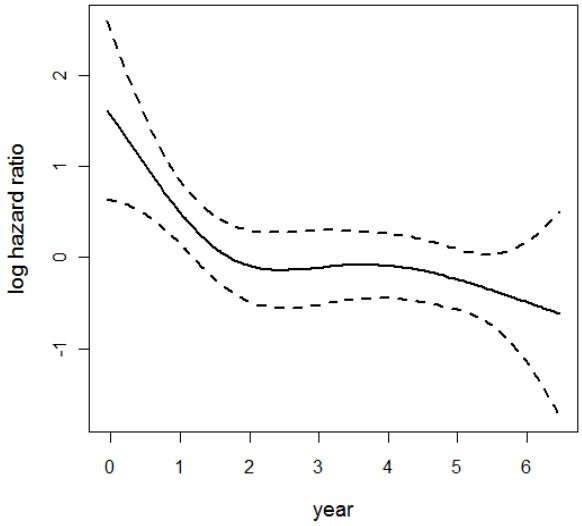


Figura 4.6- Efeito não linear para a variável *year*.

Nas análises anteriores estudou-se um possível o efeito não linear das covariáveis *age* e *year* no modelo de forma isolada. Isto é, foram ajustados dois modelos: um para testar um possível efeito não linear da covariável *age*; e outro para testar um possível efeito não linear da covariável *year*. No entanto, pode também ser pertinente considerar a suavização de duas ou mais covariáveis simultaneamente. Para tal pode-se executar no R o comando abaixo. Na Tabela 4.13 apresentam-se os respectivos resultados.

```
R> fit3=coxph(Surv(time,status)~pspline(age,0)+ pspline(year,0) +factor(surgery,
levels=c(1,0)),data=heart2)
```

Variável	***Efeito***	***Estatistica de Teste***	***Graus de Liberdade***	***Valor-Prova***
age	Linear	3.67	1	0.055
	Não Linear	15.22	8.56	0.071
year	Linear	5.51	1	0.019
	Não Linear	21.40	7.44	0.004
surgery		6.77	1	0.009
Graus de liberdade dos termos = 9.6 8.4 0.9				
Likelihood-ratio test =49.3 18.9 df, p=0.000159				

Tabela 4.13 - Averiguação de efeito linear ou não linear das variáveis *age* e *year*.

Esta abordagem, em R, não é em geral uma boa estratégia. De facto, como referido anteriormente, esta técnica selecciona a suavização óptima de cada uma das covariáveis via graus de liberdade (Critério de *Akaike*). No entanto, em contextos multivariados (mais do que uma covariável) este critério pode tornar-se muito complicado. Além do mais, como salientado por Cardarso-Suárez *et al*. (2010) o critério de AIC revela uma tendência sub-penalizadora originando graus de liberdade muito elevados.

Uma abordagem alternativa passa pela utilização dos designados modelos aditivos de Cox. Estes modelos incorporar efeitos não lineares de covariáveis contínuas. Para mais detalhes ver Cardarso-Suárez *et al*. (2010) e Haitie & Tbshirani (1990). Para se aplicar a metodologia destes modelos ao caso em estudo - incluir efeitos não lineares nas covariáveis *age* e *year* de forma simultânea - utilizou-se o programa *BayesX*. A seguir apresenta-se o respectivo *source code* do programa. Detalhes relativos ao *output* podem ser encontrados no Anexo II.

```
dataset d
d.infile  times1 delta times2 time status age year surgery, maxobs=5000 using
C:\Users\Francisco\Desktop\BayesX\examples\heart2.raw

remlreg r
r.outfile = C:\Users\ Desktop\BayesX\examples\r

logopen, replace using C:\Users\ Desktop\BayesX\examples\logheart2.txt

r.regress delta = time(baseline) + age(psplinerw2)+ year(psplinerw2)+surgery, family=cox using
d

r.plotnonp 2
r.plotnonp 3
```

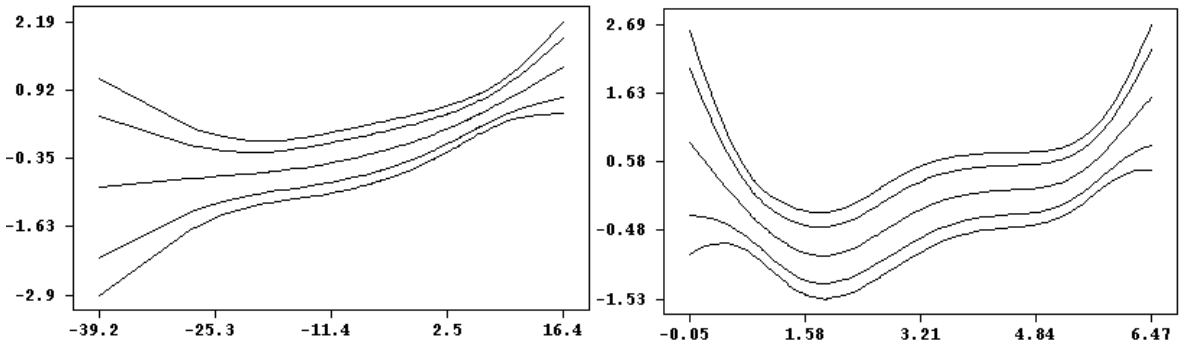


Figura 4.7 - Suavisação com intervalos de confiança pontual a 80% e 95% para as covariáveis *age* e *year* respectivamente.

A presença de efeitos não lineares pode ser inspeccionada visualmente na Figura 4.7. Adicionalmente, caso se confirme a existência de efeito não linear os respectivos gráficos são também úteis para identificar a forma correcta do referido efeito. Neste caso particular pode-se afirmar que a covariável *age* revela um efeito linear, enquanto que o gráfico da covariável *year* parece indicar um efeito quadrático (não linear). Note-se que estes resultados são concordantes com os obtidos de forma univariada, como se observa na Tabela 4.13.
De realçar, como podemos observar na Tabela 4.14, efectivamente o método AIC proporciona valores exagerados para os graus de liberdade no caso multivariado.

	age	***year***
AIC	8.56	7.44
BayesX	1.81	3.79

Tabela 4.14 – Graus de liberdade pelo método AIC versus *BayesX*.

4.3.4. Pressuposto de *Hazards* Proporcionais

A definição do modelo de regressão de Cox pressupõe que o valor de *hazard ratio* é constante. Ou seja, a razão das funções *hazard* para dois indivíduos com vectores de covariáveis fixas é constante no tempo. Assim, e estando estabelecido o modelo para estes dados, é necessário verificar a validade deste pressuposto.
O programa estatístico R disponibiliza uma função que permite verificar o pressuposto de *hazards* proporcionais descrito. Essa função é *cox.zph* e pode ser invocada conforme se indica abaixo. O *output* está organizado na Tabela 4.15.

```
R>fit=coxph(Surv(time,status)~age+pspline(year,0)+factor(surgery,levels=c(1,0)),data=heart2)
R>zph=cox.zph(fit)
R>zph
```

O valor-prova observado no GLOBAL esclarece a validade ou não do pressuposto de *hazards* proporcionais. Neste caso o valor de prova (> 0.05) releva a validade deste pressuposto. De facto, não se rejeita a hipótese de proporcionalidade (com um valor-prova muito expressivo). Os restantes valores-prova seriam apenas interessantes caso se verificasse uma violação deste

pressuposto, na medida em que permitiriam identificar qual(ais) da(s) covariável(eis) contribuia(m) para a não proporcionalidade.

	Rho	***Chisq***	***P***
Age	0.06909	0.497967	0.480
ps(year)2	-0.00503	0.000582	0.981
ps(year)3	-0.00590	0.001132	0.973
ps(year)4	-0.00652	0.001843	0.966
ps(year)5	-0.00952	0.004646	0.946
ps(year)6	-0.02199	0.027038	0.869
ps(year)7	-0.04344	0.109708	0.740
ps(year)8	-0.05295	0.165961	0.684
ps(year)9	-0.04056	0.096250	0.756
ps(year)10	-0.01081	0.006587	0.935
ps(year)11	0.01057	0.006052	0.938
ps(year)12	0.01571	0.013052	0.909
ps(year)13	0.01001	0.005226	0.942
ps(year)14	0.01123	0.006563	0.935
ps(year)15	0.02087	0.023054	0.879
ps(year)16	0.02485	0.033930	0.854
ps(year)17	0.02616	0.036337	0.849
ps(year)18	0.02654	0.032221	0.858
factor(surgery,levels = c(1, 0)) 0	-0.01773	0.024928	0.875
GLOBAL	NA	2.913.325	1.000

Tabela 4.15 - O pressuposto de *hazards* proporcionais.

4.5. Análise de Dados em Modelos Multiestado

No contexto da modelação multiestado, podemos considerar a variável transplante (*delta)* como uma covariável como um estado de risco associado, e, em seguida, usar o modelo de *illness-death* apresentado na Figura 4.8. Neste modelo, 'vivo sem transplante' indica que o indivíduo não foi sujeito a transplante, enquanto ‘vivo com transplante’ representa ter tido um transplante. Os dois estados transitórios de ‘vivo sem transplante’ e ‘vivo com transplante’ só se aplicam aos indivíduos que estão vivos, e a transição de ‘vivo com transplante’ a é assumida irreversível. ‘Morte’ é um estado absorvente que pode ser alcançado a partir de ambos os estados ‘vivo sem transplante’ e ‘vivo com transplante’.

As covariáveis *transplant* e *delta* presentes nas *packages tdc.msm* e *p3state.msm,* respectivamente, podem ser vistas como um estado de risco associado. Assim, utiliza-se o

modelo *illness-death* progressivo adaptado aos dados, apresentado na Figura 4.8. A transição $1 \rightarrow 2$ é irreversível, ou seja, não é possível transitar do estado 2 para o estado 1, devido ao facto de um indivíduo sujeito a transplante do coração, se encontrar numa condição clínica considerada como irreversível.

Com esta formulação multiestado dos dados de Stanford, estamos particularmente interessados em estudar o efeito do transplante na sobrevivência (comparando as intensidades de transição $\alpha_{12}(\cdot)$ e $\alpha_{23}(\cdot)$), assim como explorar o efeito das diferentes covariáveis em cada uma das transições.

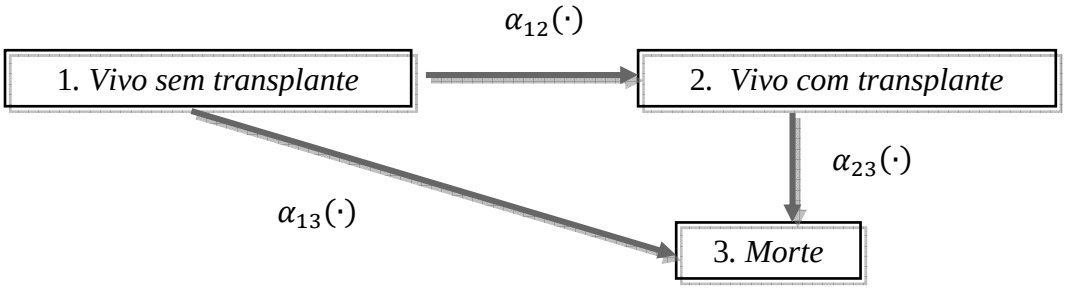


Figura 4.8 - Esquema do modelo *illness-death* para os dados de transplante de coração de Stanford

4.5.1. Modelo Tipo-Cox de Markov

A Tabela 4.16 mostra os resultados alcançados na utilização do modelo de tipo – Cox em modelos de Markov. Este modelo permite-nos especular como os efeitos das covariáveis se comportam quando as intensidades de transição são modeladas separadamente. Recorrendo à *package p3state.msm* e atendendo às variáveis que constam na base de dados *heart2*, executa-se o código seguinte no R.

```
R>library(p3state.msm)
R>data(heart2)
R>obj1.p3state<-p3state(heart2,formula=~age+year+surgery)
R>summary(obj1.p3state, model = "CMM")
```

Estados	*Estimativa*	*age*	*year*	*surgery*	*AIC*	*p-value*
$1 \rightarrow 2$	$\hat{\beta}(SE)$	0.03111 (0.0140)	0.00075 (0.0695)	0.04734 (0.3152)	509.5638	0.1234
	p-value	0.026	0.990	0.880		
$1 \rightarrow 3$	$\hat{\beta}(SE)$	0.0198 (0.0181)	-0.2833 (0.1110)	-0.2288 (0.6361)	214.9848	0.0347
	p-value	0.270	0.011	0.720		
$2 \rightarrow 3$	$\hat{\beta}(SE)$	0.0496 (0.0214)	-0.0230 (0.0969)	-0.8165 (0.4549)	290.1922	0.0102
	p-value	0.020	0.810	0.073		

Tabela 4.16 - Modelo de tipo - Cox em modelos de Markov utilizando *p3state.msm*

Atendendo aos sucessivos valores-prova apresentados na Tabela 4.16, para um nível de significância de 5% apresenta-se as seguintes conclusões.

Para o risco de morte sem transplante, $\alpha_{13}(t)$, observamos que só se obtém um efeito significativo para a variável *year*. Conclui-se assim que o ano de aceitação revela um forte efeito sobre a sobrevivência para pacientes que não chegaram a receber tranplante.

Para a intensidade da morte após a ocorrência de transplante, $\alpha_{23}(t)$, a variável *age* evidencia um efeito significativo com um valor-prova de 0.020. Estes resultados indicam para a idade há um aumento do risco $\left(\hat{\beta} = 0.0496 > 0\right)$ mostrando que quanto maior a idade maior o risco de morte (aproximadamente $44{,}198\%$ por cada ano). Similarmente, a variável *age* revela, também, ser o melhor preditor para a intensidade da doença $\alpha_{12}(t)$, sugerindo que pacientes mais velhos têm maiores probabilidades de virem a ter o transplante. Os resultados apresentados indicam que para a covariável *age* não se obtêm valores estatisticamente significativos na transição $1 \rightarrow 3$, ou seja, não se verifica um efeito significativo da variável *age* sobre a intensidade de mortalidade em pacientes sem o transplante.

No que diz respeito a covariável *surgery*, os resultados alcançados mostram não haver efeitos significativos em nenhuma das intensidades de transição. Estes resultados indicam que o facto de o indivíduo ter sido sujeito a cirurgia prévia para introdução de *bypass*, não tem influência em nenhum dos eventos (transplante e morte).

Em suma, *age* e *year* são ambas covariáveis importantes (em determinadas transições), enquanto *surgery* não apresenta efeitos significativos.

As Figuras 4.8 e 4.9 revelam a forma funcional do logaritmo da função *hazard* para as covariáveis *age* e *year* utilizando *P-splines*. Embora as curvas apresentadas forneçam informações importantes sobre o efeito de covariáveis, a sua interpretação não é simples pois não temos um valor de referência.

As curvas que proporcionam efeitos para as covariáveis, utilizando suavização *P-spline*, podem ser obtidas para covariáveis contínuas utilizando o R. A *package p3state.msm* permite obter *output* do modelo tipo-Cox de Markov. No entanto, de forma a obter a análise gráfica referida tem-se que recorrer à package *tdc.msm*. No R executa-se os comandos

```
R>tdc.msm(stanford, formula=c(6,7,8), models= "CMM", graphcov=2)
R>tdc.msm(stanford, formula=c(6,7,8), models= "CMM", graphcov=1)
```

Este modelo permite-nos observar como se comporta o efeito de covariáveis quando as intensidades de transição são modeladas separadamente. A partir das Figuras 4.9 e 4.10, pode-se observar a forma funcional do efeito de factores prognósticos em cada uma das taxas de transição (usando *P-splines* com graus de liberdade que minimizem o critério de informação de *Akaike*). A introdução de suavização (não linear), efeitos sobre as taxas de transição, muitas vezes fornece informações importantes sobre a relação entre os factores de prognóstico e risco de doença. Permite revelar, por exemplo, a presença de algum efeito de linear. Além disso, se o 'pressuposto de efeito linear' é violado, ele pode levar a conclusões estatísticas erróneas. Nas Figuras 4.9 e 4.10 não se observa nenhum efeito significativo não linear nem para a variável *year* nem para a variável *age* em nenhuma das transições (nos gráficos o ajuste observado através da linha recta encontra-se dentro dos intervalos de confiança a 95%). Não obstante, a Figura 4.10 revela também que o efeito da variável *age* na transição da mortalidade aproxima-se de um efeito quadrático, indicando que o risco diminui até cerca de 30 anos, e depois aumenta rapidamente.

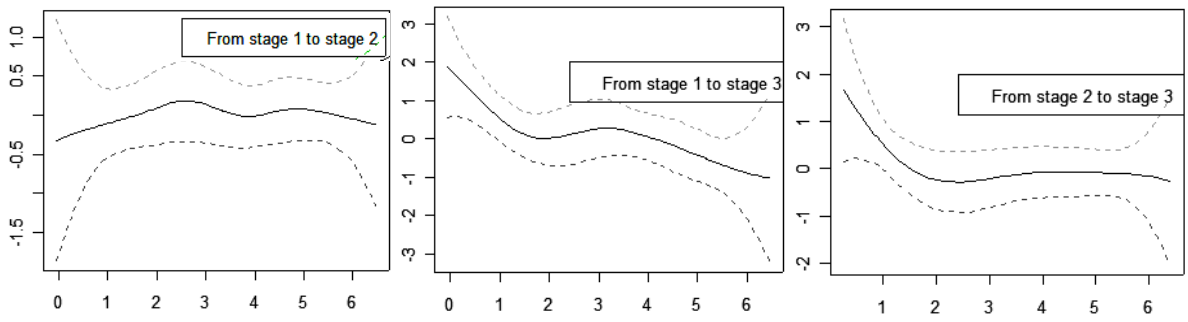


Figura 4.9 - Estimação *log hazard* com suavização *P-spline* da variável *year* (com bandas de confiança a 95%), para as transições $1 \rightarrow 2$, $1 \rightarrow 3$ e $2 \rightarrow 3$ respectivamente.

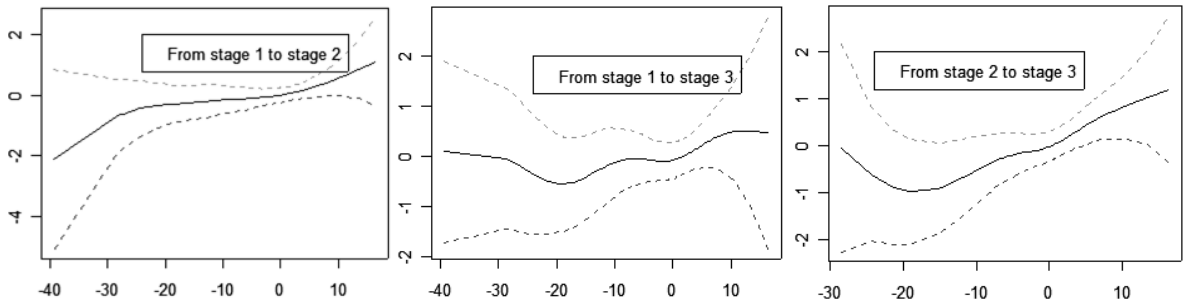


Figura 4.10 - Estimação *log hazard* com suavização *P-spline* da variável *age* (com bandas de confiança a 95%), para as transições $\mathbf{1} \rightarrow \mathbf{2}$, $\mathbf{1} \rightarrow \mathbf{3}$ e $\mathbf{2} \rightarrow \mathbf{3}$ respectivamente.

Tendo-se ajustado o modelo anterior deve-se verificar se o pressuposto de Markov é válido, isto é, se a transição futura só depende do estado no tempo *t*. Para o modelo *illness-death*, devemos testar se o tempo de permanência no estado 1 (passado) não é importante na transição do estado 2 para o estado 3. Para isso, consideramos z = tempo dispendido no estado 1, e t o tempo atual. Considere o modelo

$$\alpha_{23}(t) = \alpha_{23,0}(t)e^{\beta^T z}$$

Queremos testar se $\beta = 0$, ou seja, temos a hipótese nula, $H_0: \beta = 0$, contra a alternativa, $H_1: \beta \neq 0$.

O que equivale a testar se o modelo depende da história do processo, ou seja, se é Markoviano ou não.

Como se pode verificar no Anexo III, o efeito passado no estado 1 não é significativo ao nível de 5% (*p-value*=0.0786) não providenciando evidência contra o ajustamento do modelo de Markov aos dados.

4.5.2. Modelo de Markov em Tempo Homogéneo

Os processos multiestado são plenamente caracterizados através de intensidades de transição ou através de probabilidades de transição entre estados h e j. O modelo de Markov em tempo homogéneo proporciona uma descrição pormenorizada do processo de sobrevivência, utilizando todas as informações disponíveis para estimar o efeito de factores prognósticos e as taxas de intensidade.

Na prática, a relação entre as características individuais com as suas taxas de transição é muitas vezes de interesse, sendo o modelo de Cox de *hazard* proporcionais, uma escolha popular para modelar esta relação. As estimativas de máxima verosimilhança podem ser calculadas a partir da matriz de probabilidade de transição.

Em ambiente R, o modelo de Markov em tempo homogéneo pode ser analisado recorrendo ao *package msm* ou alternativamente ao *package tdc.msm* que oferece muitas opções e funcionalidades.

Aplicando a abordagem do modelo de Markov em tempo homogéneo, à base de dados de Stanford incluindo o potencial efeito das covariáveis *age*, *year* e *surgery* nas transições $\alpha_{12}(t)$, $\alpha_{13}(t)$ e $\alpha_{23}(t)$. Para tal, executa-se no R.

```
R>tdc.msm(stanford, formula=c(6,7,8), models= "HMM", surv.plot=T)
```

A opção HMM oferece uma descrição detalhada do processo de sobrevivência, fazendo uso de todas as informações disponíveis para estimar a probabilidade de transição e as taxas de intensidade. Em caso de HMM o teste de *Wald* pode ser aplicado para verificar diferenças nas transições para a mortalidade ou para avaliar o efeito de uma covariável considerada em duas ou mais transições.

Os resultados obtidos a partir do modelo ajustado são apresentados na Tabela 4.17, Tabela 4.18 e Tabela 4.19. no Anexo IV.

Estados	***Estimativa***	***Age***	***Year***	***Surgery***
$1 \rightarrow 2$	HR IC (95%)	1.071 (1.043 - 1.101)	1.021 (0.888 - 1.174)	1.007 (0.539 - 1.882)
$1 \rightarrow 3$	HR IC (95%)	1.051 (1.012 -1.094)	0.643 (0.450 - 0.920)	1.508 (0.349 - 6.513)
$2 \rightarrow 3$	HR IC (95%)	1.073 (1.029 - 1.120)	1.126 (0.951 - 1.336)	0.266 (0.112 - 0.631)

Tabela 4.17- Modelo de Markov em tempo homogéneo

Estimativa	*Age*	*Year*	*Surgery*
χ_1^2	0.614	6.991	7.239
$p-value$	0.735	0.030	0.027

Tabela 4.18 – Resultados do teste de *Wald* verificando as diferenças entre, respectivamente, cada uma das covariáveis (*age, year, surgery*) nas intensidades de transição.

Conforme se observa, os resultados presentes na Tabela 4.17 estão em concordância com aqueles obtidos através do modelo de Markov apresentados na Tabela 4.16. Uma excepção é o efeito da cirurgia, que não tendo apresentado um efeito estatisticamente significativo no modelo estudado anteriormente, e agora passa a ser significativa na transição $2 \rightarrow 3$. Atendendo aos resultados da Tabela 4.17 verifica-se que a idade é a única covariável a apresentar um efeito linear em todas as transições. Isso ocorre mesmo para a intensidade de transição $\alpha_{13}(t)$, algo que não ocorre quando se utiliza a abordagem de modelação de tipo-Cox de Markov. De notar ainda, que a variável *year* apresenta apenas efeito significativo para a intensidade de transição $\alpha_{13}(t)$, mostrando-se no entanto um preditor significativo.

Usamos o teste de *Wald* para verificar se existe ou não uma relação entre a ocorrência de transplante e a sobrevivência. Formalmente, a hipótese de não haver relação é dada por

$$H_0: \alpha_{13} = \alpha_{23}$$

e então o teste de *Wald* reduz-se a

$$W = \frac{(\hat{a}_{13} - \hat{a}_{13})^2}{v_{11}}$$

sendo $v_{11} = var(\hat{a}_{13} - \hat{a}_{13})$.

Sob a hipótese nula da estatística W (que segue uma distribuição χ_1^2), gera um valor de 3.121, sugerindo que o transplante conduz a uma redução no risco, mas sem atingir significância estatística ao nível de 5%, como podemos observar na Tabela 4.19.

Estimativa	$\boldsymbol{H_0: \alpha_{13} = \alpha_{23}}$
χ_1^2	3.121
$p-value$	0.077

Tabela 4.19 - Resultados do teste de *Wald* verificando se existem diferenças entre intensidades de mortalidade.

Um dos objectivos dos modelos multiestados é também estudar a evolução dos pacientes ao longo do tempo, que pode ser estudada através das probabilidades de transição. Para obter

essas estimativas recorremos à *package etm*. De modo a que a estrutura da base de danos esteja em conformidade com os parâmetros de entrada da *package etm*, é necessário efectuar uma transformação da base de dados *heart2* para o formato *heart3* (ver Tabela 4.4).

Alguns resultados para as estimativas de probabilidades de transição com o primeiro tempo igual a 0 dias ($s = 0$), utilizando a *package etm* são apresentados no Anexo V. No R executam-se os comandos a seguir apresentados (Moreira & Meira-Machado (2010)).

```
R>library(etm)
R>my.etm<-etm(data=heart3,state.names=c("1","2","3"),cens.name="cens",tra=trans,s=0)
R>p12<-trprob(my.etm,"1 2")
R>p13<-trprob(my.etm,"1 3")
```

Na Figura 4.11 apresenta-se os gráficos das probabilidade de transição para os respectivos estados. Observando a Figura 4.11 estima-se que a probabilidade de transitar para o estado 3 a partir dos estados 1 ou 2 aumenta ao longo do tempo. Relativamente à estimativa para a probabilidade transição do estado 1 para o estado 2 verifica-se que esta aumenta sensivelmente até os 65 dias e depois diminui.

Os resultados ilustrados na Figura 4.11, podem ser obtidos através do comando seguinte.

```
R>plot(my.etm,c("1 2"))
R>plot(my.etm,c("1 3"))
```

A titulo de exemplo, são os descritos no Figura 4.12 os resultados para a estimativa das probabilidades de transição para o valor inicial 90 dias ($s = 90$). A baixo apresenta-se o *input* do R.

```
R> my.etm<-etm(data=heart3,state.names=c("1","2","3"),cens.name="cens",tra=trans,s=90)
 R> p12<-trprob(my.etm,"1 2")
 R> p13<-trprob(my.etm,"1 3")
 R>p23<-trprob(my.etm,"2 3")
 R>plot(my.etm,c("1 2","1 3", "2 3"),lty=1:3, col=c("grey2","blue","green"))
```

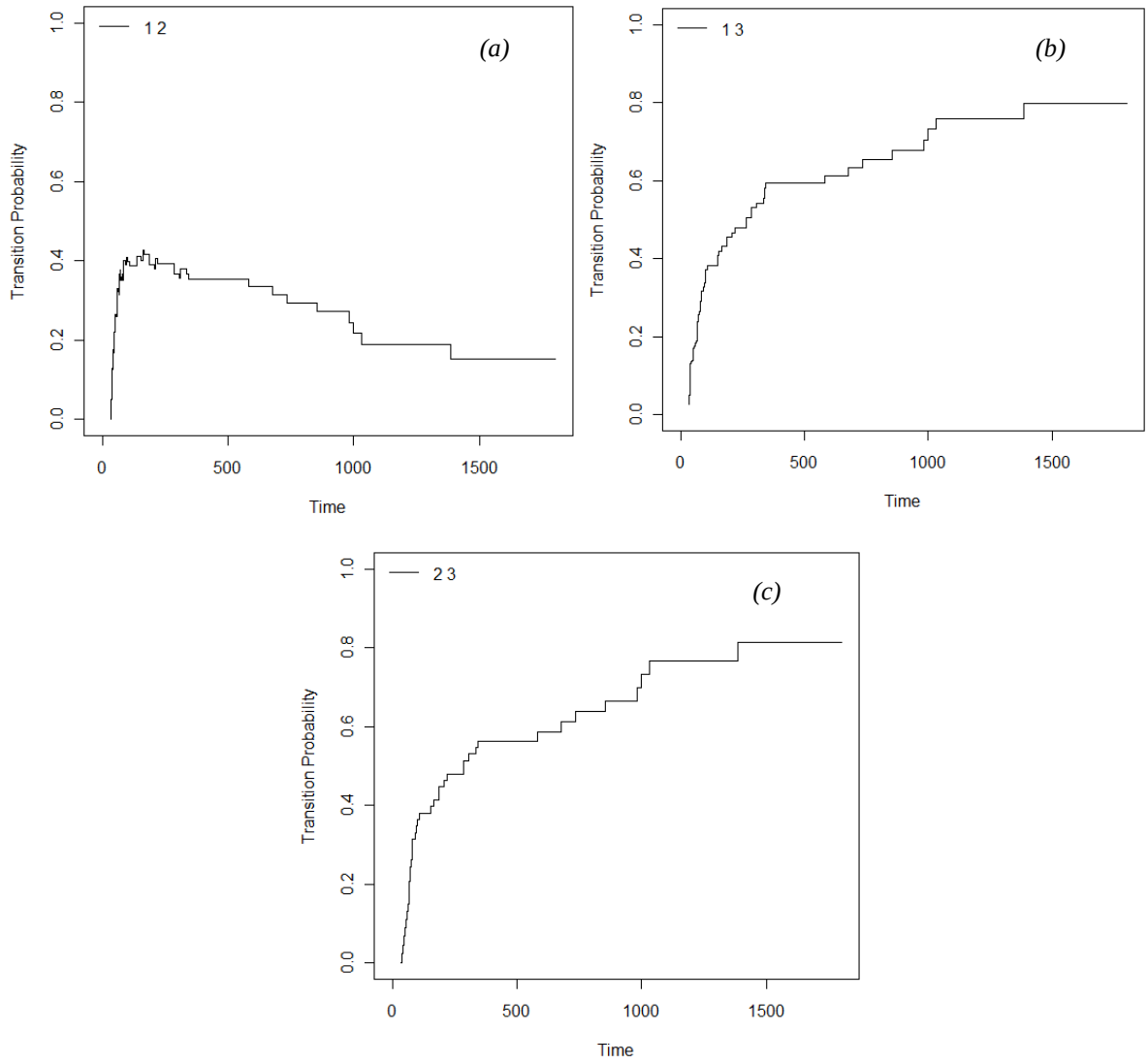


Figura 4.11 – Estimativas de probabilidades de transição tempo inicial igual a 0 dias ($\boldsymbol{s = 0}$): *(a)* do estado 1 para o estado 2; *(b)* do estado 1 para o estado 3; *(*c*)* do estado 2 para o estado 3.

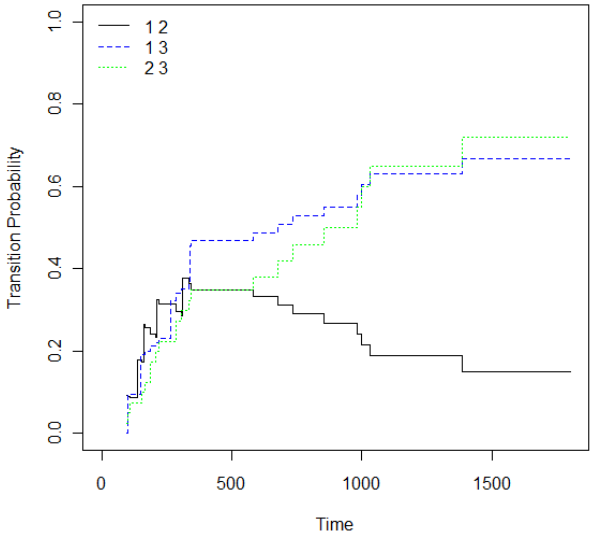


Figura 4.12 – Estimativas de probabilidades de transição com tempo inicial igual a 90 dias.

Capítulo 5

Conclusões e trabalho futuro

Nesta dissertação foram discutidos vários métodos utilizados em estudos de Análise de Sobrevivência e descrita a metodologia dos modelos multiestado, que podem ser vistos como uma generalização dos modelos de análise de sobrevivência clássica. O potencial das técnicas apresentadas foi demonstrado pela aplicação à base de dados do transplante do coração de Stanford.

Para estudar o tempo de sobrevivência dos indivíduos, estimaram-se curvas de sobrevivência segundo a proposta de Kaplan-Meier e procedeu-se ao ajustamento de modelos de regressão de Cox. Verificou-se também a validade dos pressupostos assumidos, nomeadamente o pressuposto de *hazards* proporcionais, e testou-se um possível efeito não linear para as covariáveis incluídas no modelo. Na fase de ajustamento do 'melhor' modelo válido verificou-se que os factores/covariáveis mais importantes são a idade, o ano de aceitação no estudo e a existência de cirurgia prévia para introdução de *bypass*.

No âmbito dos modelos multiestado, e em particular do modelo de *illness-death*, em que se considera a experiência de vida dos indivíduos como um processo que envolve dois estados, foram apresentados o modelo de Markov em tempo homogéneo e o modelo tipo-Cox de Markov.

Ao analisar os dados do transplante do coração de Stanford através desta metodologia, a validade do pressuposto de Markov foi analisada, tendo-se verificado que o pressuposto de Markov é satisfatório. Entre outros resultados, o modelo tipo-Cox de Markov revelou que o ano de aceitação só é importante na transição da mortalidade, sem realização de transplante.

Quanto à idade revelou efeitos significativos na transição para o transplante, e na do transplante para a morte, mas não na transição para a morte sem realização de transplante. É curioso notar que a existência de cirurgia prévia para introdução de *bypass* não apresentou efeitos significativos em nenhuma das transições.

No que concerne aos resultados do modelo de Markov em tempo homogéneo ajustado, revelaram-se em boa concordância com os obtidos através do modelo tipo-Cox de Markov, sendo as conclusões retiradas muito semelhantes. A única excepção diz respeito à covariável idade que apresentou um efeito significativo em todas as transições.

Ao destacar os factores que influenciam as diversas intensidades de transição, os modelos multiestado facultaram informações importantes, não perceptíveis no modelo de regressão de Cox ajustado considerando a experiência dos pacientes como consistindo numa única transição para a mortalidade. De facto, os resultados obtidos, usando modelos multiestado, permitiram compreender melhor o efeito factores de risco como a idade, cirurgia prévia e ano de aceitação no estudo, no desenvolvimento normal da doença.

Nesta dissertação, discutiram-se vantagens da utilização dos modelos de estados múltiplos na análise de dados de sobrevivência e ilustrou-se a aplicação destes métodos. No decorrer do trabalho realizado fizeram-se opções que determinaram o rumo seguido. Claramente outras opções poderiam ter sido tomadas, o que certamente conduziria a outro tipo de resultados e diferentes perspectivas de análise. Nesse seguimento, são feitas algumas sugestões para trabalho futuro que visam, quer completar o trabalho realizado, quer abrir novos percursos de investigação.

Assim, relativamente ao modelo multiestado ajustado aos dados do transplante do coração de Stanford seria interessante um estudo sobre as probabilidades de transição entre os estados considerados. No estudo realizado, a abordagem considerada centrou-se nas intensidades de transição, que representam o risco instantâneo de progressão entre as várias transições, e foram modeladas utilizando modelos de regressão de Cox. As probabilidades de transição permitiriam reflectir sobre a probabilidade de transição entre os diferentes estados ao longo do tempo de sobrevivência.

Outro tema que poderia ser desenvolvido no âmbito dos modelos multiestado diz respeito aos modelos Markovianos, em tempo não-homogéneo, que permitem uma generalização dos modelos usualmente empregues. Em algumas aplicações a hipótese de homogeneidade do tempo pode não ser válida, sendo recomendável a utilização de um modelo que não tenha

subjacente tal pressuposto. Dadas as contribuições metodológicas para a análise desses modelos ainda serem escassas na literatura técnica, teria interesse uma investigação sobre esses mesmos modelos.

O desenvolvimento de *source code* em ambiente R para testar um possível efeito não linear de duas ou mais covaríaveis simultaneamente, num modelo de regressão de Cox, seria outro tema a desenvolver. De facto, apesar de no estudo realizado se utilizar o software R na aplicação dos métodos apresentados, houve a necessidade de recorrer ao software *BayesX* para ultrapassar essa limitação. Deste modo, o desenvolvimento desse *source code* permitiria desenvolver todas as análises recorrendo exclusivamente ao software R.

Anexo I

Informação fornece detalhes acerca da curva de sobrevivência estimada no Capítulo 4.

t_i	n_i	$\hat{S}$	Desvio Padrao	Lim. Inf. I.C. 95%	Lim. Sup. I.C.95%	Valor -Prova
1	103	1	0.990	0.00966	0.9715	1.000
2	102	3	0.961	0.01904	0.9246	0.999
3	99	3	0.932	0.02480	0.8847	0.982
5	96	2	0.913	0.02782	0.8597	0.969
6	94	2	0.893	0.03043	0.8355	0.955
8	92	1	0.883	0.03161	0.8237	0.948
9	91	1	0.874	0.03272	0.8119	0.940
12	89	1	0.864	0.03379	0.8002	0.933
16	88	3	0.835	0.03667	0.7656	0.910
17	85	1	0.825	0.03753	0.7543	0.902
18	84	1	0.815	0.03835	0.7431	0.894
21	83	2	0.795	0.03986	0.7208	0.877
28	81	1	0.785	0.04056	0.7098	0.869
30	80	1	0.776	0.04122	0.6989	0.861
32	78	1	0.766	0.04188	0.6878	0.852
35	77	1	0.756	0.04250	0.6769	0.844
36	76	1	0.746	0.04308	0.6659	0.835
37	75	1	0.736	0.04364	0.6551	0.827
39	74	1	0.726	0.04417	0.6443	0.818
40	72	2	0.706	0.04519	0.6225	0.800
43	70	1	0.696	0.04565	0.6117	0.791
45	69	1	0.686	0.04609	0.6009	0.782
50	68	1	0.675	0.04650	0.5902	0.773
51	67	1	0.665	0.04689	0.5796	0.764
53	66	1	0.655	0.04725	0.5690	0.755
58	65	1	0.645	0.04759	0.5584	0.746
61	64	1	0.635	0.04790	0.5479	0.736
66	63	1	0.625	0.04819	0.5374	0.727

68	62	2	0.605	0.04870	0.5166	0.708
69	60	1	0.595	0.04892	0.5063	0.699
72	59	2	0.575	0.04929	0.4857	0.680
77	57	1	0.565	0.04945	0.4755	0.670
78	56	1	0.554	0.04958	0.4654	0.661
80	55	1	0.544	0.04970	0.4552	0.651
81	54	1	0.534	0.04979	0.4451	0.641
85	53	1	0.524	0.04986	0.4351	0.632
90	52	1	0.514	0.04991	0.4251	0.622
96	51	1	0.504	0.04994	0.4151	0.612
100	50	1	0.494	0.04995	0.4052	0.602
102	49	1	0.484	0.04993	0.3953	0.592
110	47	1	0.474	0.04992	0.3852	0.582
149	45	1	0.463	0.04991	0.3749	0.572
153	44	1	0.453	0.04987	0.3647	0.562
165	43	1	0.442	0.04981	0.3545	0.551
186	41	1	0.431	0.04975	0.3440	0.541
188	40	1	0.420	0.04966	0.3336	0.530
207	39	1	0.410	0.04954	0.3233	0.519
219	38	1	0.399	0.04940	0.3130	0.509
263	37	1	0.388	0.04923	0.3027	0.498
285	35	2	0.366	0.04885	0.2817	0.475
308	33	1	0.355	0.04861	0.2713	0.464
334	32	1	0.344	0.04834	0.2610	0.453
340	31	1	0.333	0.04804	0.2507	0.442
343	29	1	0.321	0.04773	0.2401	0.430
584	21	1	0.306	0.04785	0.2252	0.416
675	17	1	0.288	0.04830	0.2073	0.400
733	16	1	0.270	0.04852	0.1898	0.384
852	14	1	0.251	0.04873	0.1712	0.367
980	11	1	0.228	0.04934	0.1491	0.348
996	10	1	0.205	0.04939	0.1279	0.329
1032	9	1	0.182	0.04888	0.1078	0.308
1387	6	1	0.152	0.04928	0.0804	0.287

Anexo II

Output aplicando *BayesX* à base de dados transplante de coração de Stanford (*heart2).*

```
f_time_logbaseline


Estimated variance:  0.00590542
Smoothing parameter: 169.336
(Smoothing parameter = 1 / variance)
Degrees of freedom: 2.22469

Variance and smoothing parameter are stored in file
C:\Users\Desktop\BayesX\examples\r_f_time_logbaseline_var.res

Results are stored in file
C:\Users\Desktop\BayesX\examples\r_f_time_logbaseline.res

Postscript file is stored in file
C:\Users\.Desktop\BayesX\examples\r_f_time_logbaseline.ps

Results may be visualized using method 'plotnonp'
Type for example: objectname.plotnonp 1


f_age_pspline


Estimated variance:  0.00639782
Smoothing parameter: 156.303
(Smoothing parameter = 1 / variance)
Degrees of freedom: 1.8077

Variance and smoothing parameter are stored in file
C:\Users\Desktop\BayesX\examples\r_f_age_pspline_var.res

Results are stored in file
C:\Users\.Desktop\BayesX\examples\r_f_age_pspline.res

Postscript file is stored in file
C:\Users\Desktop\BayesX\examples\r_f_age_pspline.ps

Results may be visualized using method 'plotnonp'
Type for example: objectname.plotnonp 2
```

f_year_pspline

```
Estimated variance:  0.0573111
Smoothing parameter: 17.4486
(Smoothing parameter = 1 / variance)
Degrees of freedom: 3.78595

Variance and smoothing parameter are stored in file
C:\Users\Desktop\BayesX\examples\r_f_year_pspline_var.res

Results are stored in file
C:\Users\Desktop\BayesX\examples\r_f_year_pspline.res

Postscript file is stored in file
C:\Users\Desktop\BayesX\examples\r_f_year_pspline.ps

Results may be visualized using method 'plotnonp'
Type for example: objectname.plotnonp 3
```

FixedEffects

Variable	Post. Mode	Std. Dev.	p-value	95% Confidence Interval	
const	-5.80906	0.293799	1.1755le-14	-6.38502	-5.2331
surgery	-0.601423	0.360402	0.0947708	-1.30795	0.105106

```
Results for fixed effects are also stored in file
C:\Users\Francisco\Desktop\BayesX\examples\r_FixedEffects.res
```

Model Fit

```
-2*log-likelihood:                951.983
Degrees of freedom:               9.81833
(conditional) AIC:                971.62
(conditional) BIC:                997.488

Results on the model fit are stored in file
C:\Users\Francisco\Desktop\BayesX\examples\r_modelfit.raw
```

Anexo III

Output aplicando p3state.msm para o modelo tipo Cox de Markov, à base de dados transplante de coração de Stanford (*heart2)*.

```
Number of individuals experiencing the intermediate event:  69
Number of events for the direct transition from state 1 to state 3:  30
Number of individuals remaining in state 1:  4
Number of events on transition from state 2:  45
Number of censored observations on transition from state 2:  24


********************** COX MARKOV MODEL **********************

    *************** FROM STATE 1 TO STATE 3 ***************

n=  103
              coef exp(coef)   se(coef)          z   Pr(>|z|)
age      0.01978539 1.0199824 0.01807908  1.0943806 0.27378810
year    -0.28331015 0.7532861 0.11096315 -2.5531913 0.01067409
surgery -0.22875449 0.7955238 0.63608541 -0.3596286 0.71912491

        exp(coef) exp(-coef) lower .95 upper .95
age     1.0199824   0.980409 0.9844729 1.0567728
year    0.7532861   1.327517 0.6060493 0.9362934
surgery 0.7955238   1.257033 0.2286737 2.7675156

Likelihood ratio test=  8.623363 on  3  df, p= 0.03474115

-2*Log-likelihood= 214.9848

    *************** FROM STATE 1 TO STATE 2 ***************

n=  103
               coef exp(coef)   se(coef)          z   Pr(>|z|)
age     0.0311147186  1.031604 0.01398119 2.22546929 0.02604975
year    0.0007505999  1.000751 0.06948591 0.01080219 0.99138127
surgery 0.0473360792  1.048474 0.31524102 0.15015838 0.88063966
```

```
        exp(coef) exp(-coef) lower .95 upper .95
age      1.031604  0.9693644 1.0037190  1.060263
year     1.000751  0.9992497 0.8733322  1.146760
surgery  1.048474  0.9537668 0.5652286  1.944874

Likelihood ratio test=  5.768582 on  3  df, p= 0.1234284

-2*Log-likelihood= 509.5638

    *************** FROM STATE 2 TO STATE 3 ****************

n=  69
               coef exp(coef)   se(coef)          z   Pr(>|z|)
age       0.04956295 1.0508117 0.02137741  2.3184737 0.02042359
year     -0.02303487 0.9772284 0.09693819 -0.2376243 0.81217248
surgery -0.81647952 0.4419849 0.45491690 -1.7947883 0.07268744

        exp(coef) exp(-coef) lower .95 upper .95
age     1.0508117  0.9516452 1.0076935  1.095775
yfrom state 2 to state 3022 0.8081317  1.181708
surgery 0.4419849  2.2625206 0.1812097  1.078036

Likelihood ratio test=  11.30435 on  3  df, p= 0.01018901

-2*Log-likelihood= 290.1922

Checking the Markov assumption:
Testing if the time spent in state 1 (start) is important on transition

              coef exp(coef)    se(coef)         z   Pr(>|z|)
start -0.009392569 0.9906514 0.005340591 -1.758713 0.07862619

The p-value is  0.07862619
```

Anexo IV

Output aplicando *tdc.msm* para o modelo de Markov em tempo homogéneo, à base de dados transplante de coração de Stanford (*stanford)*.

```
*********************************************************************
   ***** MULTI-STATE HOMOGENEOUS MARKOV MODEL *****
*********************************************************************

***** Undergoing Transitions *****
    to
from 1  2  3
   1  4 69 30
   2  0 24 45

*** convergence ***
initial  value 1770.436519
iter  10 value 1709.107533
final  value 1701.271173
converged

-2*Log-likelihood: 1701.271

***** Estimated coefficients *****
$logbaseline
         Stage 1   Stage 2   Stage 3
Stage 1        0 -4.079906 -5.540312
Stage 2        0  0.000000 -6.212897
Stage 3        0  0.000000  0.000000

$age
         Stage 1    Stage 2    Stage 3
Stage 1        0 0.06898958 0.05024326
Stage 2        0 0.00000000 0.07064008
Stage 3        0 0.00000000 0.00000000
```

```
$year
         Stage 1    Stage 2    Stage 3
Stage 1        0 0.02116241 -0.4412467
Stage 2        0 0.00000000  0.1190527
Stage 3        0 0.00000000  0.0000000

$surgery
         Stage 1     Stage 2    Stage 3
Stage 1        0 0.006854728  0.4104957
Stage 2        0 0.000000000 -1.3232050
Stage 3        0 0.000000000  0.0000000

$baseline
             Stage 1      Stage 2     Stage 3
Stage 1 -0.02083436  0.016909053 0.003925302
Stage 2  0.00000000 -0.002003426 0.002003426
Stage 3  0.00000000  0.000000000 0.000000000
$baseline
             Stage 1      Stage 2     Stage 3
Stage 1 -0.02083436  0.016909053 0.003925302
Stage 2  0.00000000 -0.002003426 0.002003426
Stage 3  0.00000000  0.000000000 0.000000000

$age
                          HR      L95      U95
Stage 1 - Stage 2 1.071425 1.042991 1.100634
Stage 1 - Stage 3 1.051527 1.010531 1.094186
Stage 2 - Stage 3 1.073195 1.028606 1.119717

$year
                          HR       L95       U95
Stage 1 - Stage 2 1.021388 0.8883874 1.1743000
Stage 1 - Stage 3 0.643234 0.4495542 0.9203562
Stage 2 - Stage 3 1.126429 0.9514686 1.3335625

$surgery
                           HR       L95       U95
Stage 1 - Stage 2 1.0068783 0.5386172 1.8822346
Stage 1 - Stage 3 1.5075650 0.3489639 6.5128570
Stage 2 - Stage 3 0.2662805 0.1123546 0.6310852
```

```
***** Estimate of the ratio of the progression rate 1-3 *****
***** into   death   to   the  corresponding  rate  2-3 *****
Estimate: 1.959295
SE: 0.7459134


**** WALDS TEST TO CHECK FOR DIFFERENCES BETWEEN MORTALITY INTENSITIES ****
LET H0 BE THE HYPOTHESIS OF NON-DIFFERENCES BETWEEN MORTALITY INTENSITIES
H0 produces a chisquare statistic of: 3.121163
with a p-value of: 0.07728162


**** WALDS TEST TO CHECK FOR DIFFERENCES BETWEEN COVARIATE  1  IN THE TRANSITION INTENSITIES ****
LET H0 BE THE HYPOTHESIS OF NON-DIFFERENCES BETWEEN COVARIATE  1  IN THE TRANSITION INTENSITIES
H0 produces a chi-square statistic of: 0.6144772
with a p-value of: 0.7354751


**** WALDS TEST TO CHECK FOR DIFFERENCES BETWEEN COVARIATE  2  IN THE TRANSITION INTENSITIES ****
LET H0 BE THE HYPOTHESIS OF NON-DIFFERENCES BETWEEN COVARIATE  2  IN THE TRANSITION INTENSITIES
H0 produces a chi-square statistic of: 6.991196
with a p-value of: 0.0303306


**** WALDS TEST TO CHECK FOR DIFFERENCES BETWEEN COVARIATE  3  IN THE TRANSITION INTENSITIES ****
LET H0 BE THE HYPOTHESIS OF NON-DIFFERENCES BETWEEN COVARIATE  3  IN THE TRANSITION INTENSITIES
H0 produces a chi-square statistic of: 7.238525
with a p-value of: 0.02680244


***** Sojourn times *****
$estimate
  Stage 1   Stage 2
 47.99765 499.14494
```

Anexo V

Estimativas de probabilidades de transição tempo inicial igual a 0 dia ($s = 0$).

t	*From 1 to 2*	*From 2 to 3*	*From 1 to 3*	*t*	*From 1 to 2*	*From 2 to 3*	*From 1 to 3*
1	0.01941748	0.0000000	0.009708738	**36**	0.39303344	0.2645786	0.254428992
2	0.04854369	0.0000000	0.038834951	**37**	0.41317844	0.2645786	0.264501494
3	0.07766990	0.0000000	0.067961165	**38**	0.42325095	0.2645786	0.264501494
4	0.09708738	0.0000000	0.067961165	**39**	0.41340790	0.2816814	0.274344539
4.5	0.10679612	0.0000000	0.067961165	**40**	0.41340790	0.2816814	0.294489543
5	0.11650485	0.0909091	0.087378641	**41**	0.43355290	0.2816814	0.294489543
6	0.12621359	0.0909091	0.106796117	**43**	0.42347028	0.2983865	0.304572169
8	0.14563107	0.0909091	0.116504854	**45**	0.41338765	0.3150916	0.314654794
9	0.14563107	0.0909091	0.126213592	**46**	0.43353266	0.3150916	0.314654794
10	0.16504854	0.0909091	0.126213592	**50**	0.43353266	0.3150916	0.324727296
11	0.16504854	0.0909091	0.126213592	**51**	0.44359551	0.3310197	0.334809451
12	0.19457929	0.0909091	0.136057174	**53**	0.43351379	0.3462238	0.344891167
13	0.20442287	0.0909091	0.136057174	**57**	0.44358629	0.3462238	0.344891167
16	0.19479761	0.1774892	0.165369600	**58**	0.44357729	0.3610823	0.354972674
17	0.20474489	0.2186147	0.175109480	**60**	0.45364979	0.3610823	0.354972674
18	0.21458847	0.2186147	0.184953062	**61**	0.44356868	0.3752805	0.365053780
19	0.22443205	0.2186147	0.184953062	**66**	0.43348758	0.3894786	0.375134887
20	0.23427563	0.2186147	0.184953062	**67**	0.45363258	0.3894786	0.375134887
21	0.26380638	0.2186147	0.204640224	**68**	0.43347113	0.4166129	0.395296335
23	0.27364996	0.2186147	0.204640224	**69**	0.43347113	0.4166129	0.405368837
25	0.28349354	0.2186147	0.204640224	**71**	0.44354363	0.4166129	0.405368837
26	0.30318070	0.2186147	0.204640224	**72**	0.42338256	0.4431305	0.425529911
27	0.32286787	0.2186147	0.204640224	**77**	0.41330202	0.4563893	0.445690985
28	0.33277115	0.2422931	0.214424099	**78**	0.41329399	0.4696481	0.455771326
30	0.32298377	0.2645786	0.224211486	**80**	0.40321365	0.4825835	0.465851668
31	0.34267093	0.2645786	0.224211486	**81**	0.39313331	0.4955189	0.465851668
32	0.36281593	0.2645786	0.234283988	**83**	0.41327831	0.4955189	0.475924169
33	0.37288844	0.2645786	0.234283988	**85**	0.41327831	0.4955189	0.486004128
35	0.37288844	0.2645786	0.244356490	**90**	0.40319835	0.5078234	0.496084087

96	0.40319089	0.5201278	0.506163859	**428**	0.30174911	0.6789048	0.678105891
100	0.39311112	0.5321246	0.516236361	**445**	0.30174911	0.6789048	0.678105891
102	0.39311112	0.5321246	0.516236361	**482**	0.30174911	0.6789048	0.678105891
109	0.39311112	0.5321246	0.526581391	**515**	0.30174911	0.6789048	0.678105891
110	0.38276609	0.5444371	0.526581391	**545**	0.30174911	0.6789048	0.693193346
131	0.38276609	0.5444371	0.526581391	**584**	0.28666165	0.6949595	0.693193346
139	0.39283859	0.5444371	0.536653893	**596**	0.28666165	0.6949595	0.693193346
149	0.39283859	0.5444371	0.547271152	**620**	0.28666165	0.6949595	0.693193346
153	0.38222133	0.5567496	0.547271152	**670**	0.28666165	0.6949595	0.711109699
160	0.39229384	0.5567496	0.557873688	**675**	0.26874530	0.7140246	0.729026052
165	0.38169130	0.5687293	0.557873688	**733**	0.25082894	0.7330896	0.729026052
180	0.38169130	0.5687293	0.568779154	**842**	0.25082894	0.7330896	0.748320586
186	0.37078583	0.5810514	0.579684620	**852**	0.23153441	0.7536212	0.748320586
188	0.35988037	0.5933734	0.590590085	**916**	0.23153441	0.7536212	0.748320586
207	0.34897490	0.6056954	0.590590085	**942**	0.23153441	0.7536212	0.771474027
210	0.35904740	0.6056954	0.601470310	**980**	0.20838097	0.7782591	0.794627468
219	0.34816718	0.6176440	0.611542812	**996**	0.18522753	0.8028969	0.817780909
263	0.34816718	0.6176440	0.611542812	**1032**	0.16207409	0.8275348	0.817780909
265	0.34816718	0.6176440	0.634005210	**1142**	0.16207409	0.8275348	0.817780909
285	0.32570478	0.6423121	0.645236410	**1322**	0.16207409	0.8275348	0.850195727
308	0.31447358	0.6546462	0.645236410	**1387**	0.12965927	0.8620279	0.850195727
310	0.32454608	0.6546462	0.656427654	**1401**	0.12965927	0.8620279	0.850195727
334	0.31335484	0.6665550	0.666500156	**1408**	0.12965927	0.8620279	0.850195727
340	0.31335484	0.6665550	0.678105891	**1572**	0.12965927	0.8620279	0.850195727
343	0.30174911	0.6789048	0.678105891	**1587**	0.12965927	0.8620279	0.850195727
370	0.30174911	0.6789048	0.678105891	**1800**	0.12965927	0.8620279	0.850195727
397	0.30174911	0.6789048	0.678105891				

Bibliografia

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, **6**, 701-726.

Akaike, H. (1974). A new look at the statistical model indentification. *IEEE Transactions on Automatic Control*, **19** , 716-723.

Andersen, P. K., & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research* , **11**, 91-115.

Andersen, P. K., Borgan, O., Grill, R., & Keiding, N. (1993). *Statistical Models Based on Counting.* New -York: Springer - Verlag.

Anderson, G. L. & Fleming, T. R. (1995) Model misspecification in proportional hazards regression. *Biometrika*, **82**, 527-541.

Athanasiou, T., & Darzi, A. (2010). *Key Topics in Surgical Research and Methodology*. Springer.

Bahrawar, J. (2005). *Improved Inferences in the Context of Survival/Failure Time.* Peshawar: PhD thesis, University of Peshawar, Peshawar.

Belitz, C., Brezger, A., Kneib, T., & Lang, S. (2009). BayesX - Software for Bayesian inference in structured additive regression models. Version 2.0.1. *disponível em http://www.stat.uni-muenchen.de/~bayesx* .

Brezger, A., Kneib, T. & Lang, S. (2005). BayesX: analyzing bayesian structured additive regression models. *Journal of Statistical Software*, **14**, 1–22 (http://www.jstatsoft.org/).

Borgan, Ø. (1997). *Three contributions to the Encyclopedia of Biostatistics:The Nelson Aalen, Kaplan Meier, and Aalen Johansen estimators*. Tech. Rep., Department of Mathematics, University of Oslo.

Botelho, F., Silva, C., & Cruz, F. (2009). Epidemiologia explicada - Análise de Sobrevivência. *Actas Urológia , **26*** , 33-38.

Breslow, N. & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics*, **2**, 437-453

Cadarso-Suárez, C., Meira-Machado L. & Gude, F. (2010) Flexible hazard ratio curves for continuous predictors in multi-state models: a P-spline approach, *Statistical Modelling,**10 (3)**, 291-314.*

Commenges, D. (1999). Multi-state Models in Epidemiology. *Lifetime Data Analysis, **5*** , 315–327 .

Cox, D. R., & Oakes, D. (1984). *Analysis of Survival Data.* London, N Chapman and Hall, London – New York 1984.

Cox, D. R. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society. Series B* **34** (2), 187–220.

Cox, D. R. (1975). Partial Likelihood. *Biometrika,* **62(2)**, 269-276

Crowley, J., & Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association,* **72** , 27–36.

Ettore Marubini, M. G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies.* Vic Barnett.

Ferreira, J. M. (2007). *Anáslise de Sobrevivência: uma Visão de Risco Comportamental na Utilizaçao de Cartão de Crédito*. Dissertação (Mestrado em Biometria e Estatística Aplicada) - Universidade Federal Rural de Pernambuco.

Fleming, T. H. & Harrington, D.P. (1984). Nonparametric estimation of the survival distribution in censored data. *Comm. in Statistics*, **13**, 2469-86.

Fleming, T. R. &Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, New York: Wiley.

Frydman, H. (1995). Nonparametric estimation of a Markov illness-death process from interval-censored observations, with application to diabetes survival data. *Biometrika,* **82**, 773-789.

Greenwood, M. (1926). Reports on Public Health and Medical Subjects. London: *Her Majesty's Stationery Office*, **33**, 1–26.

Hair, J., Tatham, R., Anderson, R., & Black, W. (1998). *Multivariate Data Analysis, 5 ed.* New Jersey: Prentice-Hall, Inc.

Hall, W.J. & Wellner, J.A. (1980). Confidence bands for a survival curve from censored data. *Biometrika*, **67**,133- 143.

Hastie T. J. & Tibshirani R.J. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, **46**, 1005-1016.

Hougaard, P. (2000). *Analysis of multivariate survival data.* New York, Berlin,Heidelberg, Springer.

Hougaard, P. (1999). Multi-state Models: A Review. *Life Data Analysis* , **5**, pp. 239–264.

Hurtado, H. (2008). *Análisis de supervivencia en fiabilidad. Predicción en condiciones de alta censura y truncamiento: el caso de las redes de suministro de agua potable.* Valência: Universitat Politécnica de Valência.

Johansen, S., & Aalen, O. (1978). An empirical transition matrix for nonhomogeneous Markov. *Scandinavian Journal of Statistics* , **5**, 141-150.

Kalbfleisch, & Prentice. (1980). *The Statistical Analysis of Failure Time Data*, New. York: John Wiley & Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

Klein, J. P. (1991). Small sample moments of some estimators of the variance of the Kaplan-Meier and. Nelson-Aalen estimators. *Scand. J. Statist*, **18**, 333-40

Mason R. L., Gunst, R. F. & Hess, J. L. (2003). *Statistical Design and Analysis of Experiments: With Applications to Engineering and Science*. J. Wiley, New York

Mantel, N., & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *J*ournal of the National Cancer Institute, **22**, 719-748

Marubini, E. & Valsecchi, M.G (1995). *Analysing Survival Data from Clinical Trials and Observational Studies. Wiley*, Chichester, England.

Meier, P. (1975). Estimation of a distribution function from incomplete observations. *In Perspectives in Probability and Statistics, Ed. J. Gani*, 67- 87.

Meira-Machado, L., Cadarso-Suárez, C. & de Uña-Álvarez, J. (2008) Inference in the progressive three-state model. *Intern. J. Mathematical Models and Methods in Applied Sciences* 3, **2**, 447-454.

Meira-Machado, L. (2006). Análise de dados de cancro do intestino utilizando modelos de multiestado, Departamento de Matemática para a Ciência e Tecnologia, Universidade do Minho.

Meira-Machado, L. (2002). *Tese de mestrado: Una revision sobre la suavización en el modelo de azares proporcionales.* Santiago de Compostela.

Meira-Machado, L, Roca-Pardinas, J (*2011*). p3state.msm: Analyzing Survival Data from an Illness-. Death Model. *Journal of Statistical Software,* **38**(3), 1-18

Meira-Machado, L,, de Uña-Álvarez, J., Cadarso-Suárez, C. & Andersen, PK. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research* **18**, 195-222

Meira-Machado, L., Cadarso-Suárez, C., & Uña-Álvarez, J. (2007) tdc.surv: An R library for the analysis of multi-state survival data. Computer Methods and Programs in Biomedicine, **865,** 131-140.

Meira-Machado, L., de Uña-Álvarez, J. & Cadarso-Suárez, C. (2006) Nonparametric estimation of transition probabilities in a non-Markov illness-death model, *Lifetime Data Analysis*, **12**, 325-344.

Moreira, A. &, Meira-Machado, L. (2010). Available Software for the Analysis of Multi-State Survival Data. XXXII Congreso Nacional de Estadistica e Investigación Operativa.

Murthy, D. N., Xie, M., & Jiang, R. (2004). *Weibull models.* New Jersey: John Wiley & Sons.

Nelson, W.B. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*,**14**, 945-965.

Oliveira, J. G., Cruz, F. R., & Colosimo, E. A. (2007). Avaliação e correção de viés no modelo de regressão de cox . *I ERPO NE* .

Poulsen, G. N. (2006). *Application of multi-state models in a study of diabetic nephropathy.* Master's thesis, Department of Mathematical Sciences, University of Copenhagen.

Rafael Pérez-Ocón & J. E. (2001). Non-homogeneous Markov models in the analysis of survival after breast cancer. *Jornal of the royal statistical society* , **50**, 111-124.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika,* **69** , 239-41.

Segão, S. (2000). Estudo Comparativo de Vários Métodos Para A Estimação de um Modelo de Sobrevivência Com Uma Covariável Mal Classificada,Departamento de Matemática, Instituto Superior Técnico, Universidade Técnica de Lisboa.

Struthers, C., & Kalbfleisch, J. (1986). Misspecified Proportional Hazards Model. *Biometrika, **74*** , 363-369.

Venables, W. & Ripley, B. D. (1999). *Modern applied statistics with S.* Springer-Verlag.

Wilson, A., Armijo, Limnios, N. & Keller-McNulty, S. (2005). Modern statistical and mathematical methods in reliability. Edited volume from the 2004 international symposium on Mathematical Methods in Reliability, World Scientific Press.

Tibshirani R, Hastie T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82**, 559-567.

Tsiatis, A. (1981). A Large Sample Study of Cox's Regression Model. *Annals of Statistics, **9***, 93-108.